Multimedia Pivot Tables for Multimedia Analytics on Image Collections

Marcel Worring *Senior Member, IEEE* , Dennis Koelma, Jan Zahálka

***Abstract*—We propose a multimedia analytics solution for getting insight in image collections by extending the powerful analytic capabilities of pivot tables, found in the ubiquitous spreadsheets, to multimedia. We formalize the concept of multimedia pivot tables and give design rules and methods for the multimodal summarization, structuring, and browsing of the collection based on these tables, all optimized to support an analyst in getting structural and conclusive insights. Our proposed solution provides truly interactive analytics on the visual content of image collections through concept detection results, as well as tags, geolocation, time and other metadata. We have performed user experiments with novice users on a dataset from Flickr to improve the initial design and with expert users in marketing and multimedia analysis on two domain specific datasets collected from Instagram. The results show that analysts are indeed capable of deriving structural and conclusive insights using the proposed multimedia analytics solution. On our website videos of the system in action are available.** [1]

***Index Terms*—Exploration; Visual Analytics; Information Visualization; Insight;**

I. Introduction

Visual collections contain a wealth of information for analytic purposes. For media companies, news agencies, and marketing managers alike social media is a crucial resource in which visual data plays an ever increasing role. In social science, biology, astrophysics, or medicine, images are a valuable source of scientific knowledge. Visual information is also becoming a prime carrier of evidence and clues in forensics and security. In all these application domains the deluge of visual collections holds tremendous value, but how do we gain insight in such a collection?

The field of visual analytics [21][39] addresses insight gain in an increasing number of scientific and applied domains. Mature visual analytics techniques exist for various kinds of data, ranging from classic factual datasets to specialized data types such as temporal or geospatial data. An example of a commercial system for factual and geospatial data is Tableau, building upon the query paradigm in [38]. Yet multimedia analytics, the combination of multimedia analysis and visual analytics [6] is just starting as a field and its potential still needs to be unlocked. Multimedia insight might pertain to a lot of different elements of the dataset. For example, it might be found in the content of the images in relation to the tags and their geographic distribution. Or it requires to study what scenes are depicted on images for different quality ratings by clients and their dependence on the age of the client making the judgment. Multimedia analytics tasks have been placed on an exploration-search axis, with the analyst tilting between the two extremes as she progresses towards insight [47]. To successfully guide the analyst's quest for insight in these diverse tasks, multimedia analytics needs to go beyond direct application of existing visual analytics techniques.

Multimedia analytics faces a number of challenges. The first one is the specific nature of images. A brief look at a small number of images is sufficient to reveal their meaning and relations. Based on this we immediately trigger knowledge we have about these objects [44]. Truly analyzing images requires seeing each image at sufficient resolution, which brings a high cost of screen space in the visualization. Moreover, two challenging gaps need to be taken into account: the semantic gap and the pragmatic gap. The semantic gap postulates that the human's ability to analyze semantic content in multimedia and put it in context dramatically surpasses that of a machine [36]. Although recent progress in deep learning [26] has significantly reduced the gap, the capability of experts assessing single images is still beyond system performance. More importantly, the pragmatic gap which conceptualizes the difference between the flexible analytic categorization in the user's mind and the rigid, pre-defined categorization performed by the machine when modeling user intent [47] is seldom addressed in the multimedia analysis literature. As a result, we cannot rely on automatic analysis alone for insight gain. In order to tackle multimedia analytics challenges, the approaches have to fully support the analyst's exploration of all information channels while aiding the analyst with machine data processing techniques.

Analytics tools are many and each of them has the potential to be adapted in such a way that it can incorporate visual data and multimedia analysis. Due to their ubiquitous presence and their proven power in a broad range of applications, we propose to consider spreadsheets as candidate analytic tools to extend to multimedia. Some interesting early approaches in this direction, solely relying on the metadata for querying and visualizing the image collection, have been proposed in [5][20]. The MediaTable [9] adds automatic concept detection to analyze and describe the visual content and organizes the results in the tabular form akin to what is used in spreadsheets. Arguably one of the most powerful features of spreadsheets for understanding complex data is the pivot table. These tables let the user interactively create summaries of the data in the spreadsheet in various ways. In this manner the user

The authors are with the Informatics Institute of the University of Amsterdam, m.worring@uva.nl, d.c.koelma@uva.nl,j.zahalka@uva.nl

[1] https://staff.fnwi.uva.nl/m.worring/pivot-tables.html

gets different views of the data, aggregated along different dimensions, and hence discovers patterns and trends. But pivot tables are primarily based on nominal, ordinal, and numeric variables. If we could have a paradigm similar to pivot tables for all the richness of an image collection containing images, tags, geographic data and any other variables it would give a truly multimodal summarization tool with different aggregated views on the data all in the hands of the user.

Multimodal summaries are powerful, yet getting insight into a complex multimedia collection is an iterative process composed of multiple steps in which the results of several summaries need to be combined. During the process the insight builds up [31] and a true multimedia analytics solution should take the characteristics of insight into account to support the user with tools addressing this buildup. How to design multimodal summarization methods for image collections with the same power as pivot tables and how to embed them in the analytics process of building insight are open questions, which we aim to address here.

In this paper, extending upon our conference paper [45], we make the following contributions:

- Formally define the concept of multimedia pivot tables.
- Intricately link multimedia analytics with multimedia pivot tables to the characteristics of insight.
- Develop design rules for multimedia pivot tables and analytics based on the characteristics of the data and insight.
- Provide a user study with novice and expert users.

II. Related work

We consider related work along three lines. First we consider methods which are also addressing large image collections ranging from visualization only to multimedia analytics following various paradigms. We then consider methods aiming at aggregating and visualizing text and metadata along various dimensions using metaphors other than a spreadsheet. Finally, we review methods which are following the spreadsheet paradigm.

A. Multimedia Visualization and Analytics

As search engines are still the primary way to access image collections visualizing a set of ranked images in a 2D grid, using implicit reading order, is still dominant. Such a grid doesn't reflect the inherent uncertainty in the content analysis of each image. In [43] they therefore organize the query result as a spiral starting in the center with size reflecting the analysis score. For longer lists [48] proposes to decompose the list into several linked grids, each organized to reflect similarity instead of ranking. Photoland [33] combines grids with groupings based on temporal and spatial information. The grid based methods make optimal use of available screen estate, yet for relations they can only show crude approximations.

Similarity based methods start from the relations among all images instead of a ranking. Such methods project the data from the high dimensional space induced by all (visual) similarities among images into the two dimensional screen space. The main goal of the projection techniques is to best preserve in the visualization the relations among the images in the high dimensional space. This leads to many overlapped images, [30] therefore adds overview and visibility as criteria to optimize. To allow for local interaction [18] provides efficient local projections. MediaGlow [13] combines similarity based visualization with additional metadata based visualizations. The difference in metadata values can also be used as a similarity function so the projection methods might create groups of images sharing similar metadata. The method in [46] combines an MDS based 2D similarity based visualization with a rank based visualization based on query by example. Finally, [34] considers color similarity in a 3D image browser. These similarity based visualizations are good for showing relations among the images while showing relations between the images and the metadata is missing or difficult.

Methods for video summarization are many (see [41] for a good overview). More recently [23][24] utilize image data to summarize videos in an optimal way. These methods are targeting automatic summarization. A good overview of video interaction methods is provided in [35], but they do not consider the visual analytics part of video. The method in [4] performs visual analytics on large scale video data by analyzing and visualizing clickstreams, but without analyzing the visual content of the videos.

Methods combining content-based multimedia analysis of images and metadata with visual analytics [6] are still limited. One of the first systems doing so is the Informedia system which employs speech and video analysis in conjunction with effective user interfaces [7]. A system targeting computer vision algorithm developers is presented in [16] allowing the user to gain insight in features and how to use these features in surveillance. Canopy [3] is an advanced system combining text analysis, various visualizations, and visual similarity based matching to explore visual collections. Hierarchical visualizations for understanding news data, based on analysis of closed-captions, are presented in [27] using hyperbolic trees as visualization to explore the hierarchical structure of the data. Concept detectors are used in both [46] and [9]. Where [46] uses grids and similarity based displays, [9] uses a heatmap like visualization simultaneously showing a large set of ranked lists. The above methods are all essentially search-focused. This means that while they have certainly started to unlock the potential that multimedia analytics methods can bring, there is little explicit support for tasks in the spirit of the visual analytics cycle by Keim et al. [22], which emphasizes iterative build-up of knowledge.

Emphasizing interaction, visualization mosaics [28] provide a very flexible way of creating a summary putting the control fully in the hand of the user. This allows for easy personalization and targeted presentations for different audiences. Such full flexibility also limits explicit structure or constraints on the final result which makes it more difficult to interactively build up insight. Mosaics are therefore well suited as summaries for presentation, they have less utility in analytics. Their solution also doesn't consider content analysis.

B. Visualizing and summarizing text and metadata

To visualize text and non-visual metadata there is a huge amount of advanced literature available. In recent years a number of excellent surveys have appeared on the visualization of different modalities with interactive websites supporting faceted search to select visualizations on the basis of their characteristics. Particular examples are text [25] and temporal-spatial data [1]. Most of these provide complete full-screen visualization systems targeting a specific purpose where all visualization components contribute to the task.

When summarizing data in a small space simple and effective visualizations are needed. A number of these are described in [11]. Many of these can be combined into one screen, yet by their simplicity they only give limited insight in the data and drill-down is needed to get a better understanding.

The above full-screen and small space based visualizations form the extremes of a continuum of visualizations yielding a comprise between simplicity and expressiveness, which one to choose is application dependent.

C. Spreadsheets and their extensions

Spreadsheets are among the best known analytic tools and their origins go back a long time. The tabular visualization which forms the basis for any spreadsheet is simple yet effective. Early information visualization systems already employed the important mechanism of being able to coordinate sorting of one column with all others. Through the visualizations in the different columns correlations become evident. The LineUp [14] takes the coordination a step further and also visualizes the relations among the different columns. No other spreadsheet functionality is added though.

Modern spreadsheets allow for the integration of visual data into the spreadsheet [12] and connecting it to web resources. They do not make images a true part of the analytics process. Integrating the spreadsheet paradigm with images is proposed in [5] to visualize the effect of different parameter settings in an experimental setting. The Photospread [20] system extends the spreadsheet to image collections by allowing groups of images in the individual cells with formula like definitions to fill the cells with the user desired selection. Using formulas in grid cells is a first step towards multimedia analytics using the capabilities of spreadsheets. The above methods are, however, all based on the standard table based spreadsheet.

Pivot tables are a way of summarizing data in a spreadsheet which can be interactively defined by a user. At its core the pivot table is a matrix of cells which can contain values. The power of pivot tables comes from the flexibility in assigning different roles for the variables in the dataset [17]. The Datacube [15] has introduced a powerful mechanism for organizing data supporting the pivot table summaries in spreadsheets. The flexible and powerful mechanism underlying the Tableau system is based on a language which is very similar to the Datacube mechanism integrating it with constraints on the possible visualizations [38]. The way users interact with the system bears great similarities with pivot tables. Extensions of pivot tables in the classic sense to image collections are limited. The PhotoCube system [40] makes a step in the direction. They rely, however, on a 3D space rather than a 2D pivot table. The recently demonstrated ICLIC system [42] incorporates a simple pivot model using histograms of stacked images.

Overall, the related work on spreadsheets provides many of the ingredients required for truly harnessing the power of this familiar visual metaphor in multimedia analytics. The multimedia pivot tables presented in this paper integrate many of these techniques and extend them in order to facilitate analytic insight in multimedia collections.

III. Methods

A pivot table in the classic sense is created by assigning variables to be used as *filter* to select the part of the dataset to work on, a variable to define the *rows* of the table, and one to define the *columns* of the table. The set of elements for each cell is now restricted to those having the corresponding row and column nominal label. The final variable determines the *value* to be used for the cell, which are integrated over the set of elements by applying a user selected *aggregation operator* such as mean, mode, or maximum value. Commonly pivot tables are restricted to categorical and numerical variables.

Our methodology extends the idea of pivot tables to the multimedia domain. In particular, in this section we first elaborate on the multimedia data representation we employ and from there consider how the vague notion of insight can be operationalized for our setting. Having done so we define the underlying pivot table mechanisms and analysis techniques. Finally, we consider the visualization of and interaction with the image collection in the multimedia pivot table to create a true multimedia analytics solution. An illustration of our proposed solution for multimodal summaries, building upon the pivot table paradigm, is presented in figure 1.

A. Multimedia Data Representation

To arrive at a multimedia analytics solution for image collections, we first have to consider the characteristics of the data in multimedia collections and how this affects their use in an analytics context. Items in an image collection comprise many different variables. First we have the *images* themselves. The visual concept descriptors resulting from the automatic analysis of the images form a solid data representation basis for multimedia analytics. Recent progress in the field has brought vocabularies ranging from 100 to over 10,000 visual concepts [10]. For a given dataset, or a subset thereof, a limited number of relevant concepts is usually more appropriate. In our case, the images are analyzed using visual detectors from the vocabulary following the analysis pipeline in [37]. For each visual concept it yields a *concept score* between 0 and 1 as an indication of concept presence. Each of the images might have a set of *tags*, taken from an unconstrained vocabulary, which could range from describing specific aspects of the content to personalized interpretations or context of the images. In addition to those we have the *attributes* composed of numeric, categorical, temporal, or geolocation variables describing the image and its context. It is this richness of modalities that we aim to give access to in our multimedia analytics solution. So

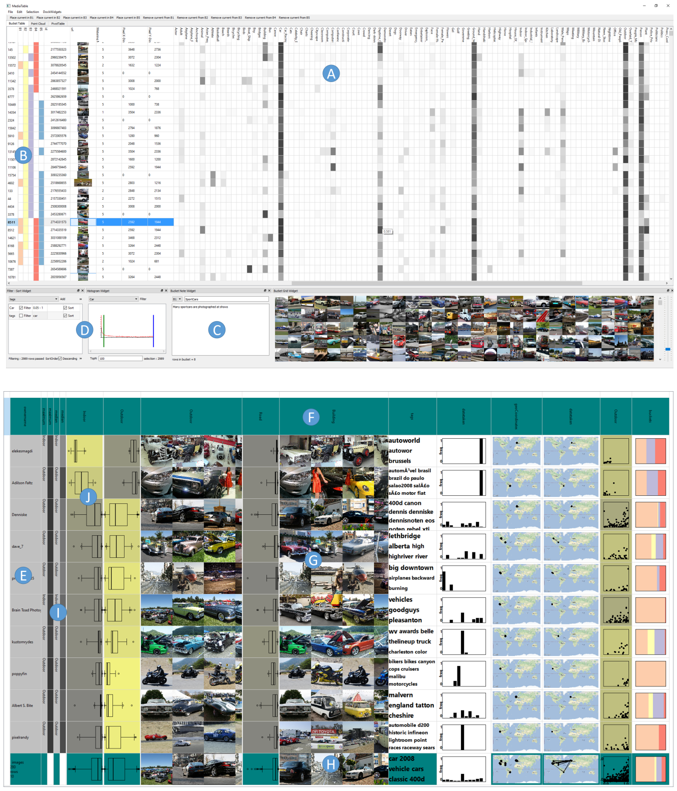


Fig. 1. An example of our multimedia pivot tables showing our faceted filtering combined with a search interface based on our Mediatable [9] to show the row based dataset (top) and the pivot table (bottom). A) the dataset with every row containing an image, its metadata and visual concept detection results. B) the current category membership of of the images and C) a widget for annotating the different categories. D) a faceted filter to determine the active set and sorting the table based on one of its columns. E) a decomposition of the dataset by a specific variable. F) sort-weight variables as columns and G) various value variables with different visualizations which are H) aggregated per column and for all visual concepts I) aggregated over individual rows. J) rows are sorted according to a relevance function based on one or more variables.

we define a multimedia image collection M in the following way.

$$M := (V^i_{\text{image}}, (V^i_{\text{type}})^*)^*_{i=1,|M|}$$

with type $\in$ {tags, text, concept, geo, temporal, nominal, ordinal, numeric}, $|\cdot|$ denoting the number of elements in a set or vector, and $(\cdot)^*$ denoting zero or more instances. Note that we made the simplifying assumption that any item in the collection is composed of exactly one image and all its associated descriptors and attributes. The definition could easily be extended to take multimedia items like text based social media posts with multiple images per item into account.

B. Multimedia insights

Insight is hard to define in a precise way, [31] however, identifies five major characteristics of insight. These characteristics can be rephrased directly into design rules as follows:

- *Complex:* Dealing with the complexity of insight requires to work with all or large amounts of the given data in a synergistic way, not simply individual data values.
- *Deep:* We need to provide the user with support to build up insight over time, letting it accumulate and build on itself to create depth, to generate further questions and, hence, further insight.
- *Qualitative:* Insight is not exact, so the representation should allow for uncertainty and subjectivity, and it should have multiple levels of resolution.
- *Unexpected:* Support should be given to get unpredictable, serendipitous, and creative insight.
- *Relevant:* The insight gained should be deeply embedded in the data domain, so we should allow to connect the data to existing domain knowledge to give it relevant meaning, going beyond dry data analysis to relevant domain impact.

The characteristics of insight dictate that multimedia analytics is an interactive process composed of multiple steps in which the user, having some high level goal, gets closer and closer to realizing it. During the process the user is employing several methodologies to access the data which can be conveniently mapped to the *exploration-search axis* [47]. Search methods for multimedia have become mature and ranking methods are many [29]. Our focus here is on the exploration part of the axis with core tasks being *summarization*, *structuring*, and *browsing*.

For structuring the collection various, not necessarily disjunct, groupings of the data form the basis. In [47] we argue, based on an extensive literature review, that for image collections these groupings or analytic categories are an essential ingredient in insight gaining processes. To be precise, when category definition is in the hand of the interacting expert the labels given by the user to categories, being it derived in a unsupervised or supervised way, provide the current view of the expert on the domain dependent relevant terms in relation to the collection. We therefore define

- *Structural insight* — a user defined label for an analytic category posing structure on the collection at the semantic level.

But the label alone is not enough to capture insight, we need to take into account the qualitative conclusions the users reach based on the analytic categories in connnection to their prior knowledge and connotations they have with the data which we define as:

- *Conclusive insight* — a qualitative conclusion about an analytics category or the data in general.

All insight is derived from the members assigned to the different analytic categories. To emphasize the action of adding items to categories we denote the categories as buckets [9], a metaphor for a place where elements can be put, each having a color consistently used throughout the system. A bucket $\mathcal{B}$ is thus defined by:

$$\mathcal{B} := (\text{label}, (\text{member})^*, (\text{conclusion})^*)$$

As a consequence of the pragmatic gap [47] discussed earlier, all the elements of $\mathcal{B}$ should be dynamic and under full control of the interacting user. The user should not only be allowed to add items to a bucket, but should have the option to reconsider membership while getting better understanding of the data. Buckets which were originally separate might be joined into one, or the items in a bucket can be redistributed over two or more sub-categories, hence creating new buckets. Adding set operations on pairs of buckets (in our system the AND and NOT operator) yields another powerful mechanism to let buckets even better capture insight. And whenever one of the above operations is applied the user will likely reflect this by changing the labels of the buckets. So we define the insight $\mathcal{B}$ in a time dependent manner (denoted $\mathcal{B}^t$) as follows:

$$\mathcal{B}^t := \{B_i^t\}_{i=1,|\mathcal{B}^t|}$$

C. Multimedia Pivot Variables

Having defined the insight in terms of labels and structure we now consider the multimedia summarization step based on pivot tables. Like in the standard pivot tables variables can take different roles and this is where the power of pivot tables emerges. Yet for multimedia data this is not as straightforward as it is for regular pivot tables. We now elaborate on these different roles and how we define them for multimedia.

The first role of a variable we consider is when it is used as a filter. As multimedia collections are large and composed of various types of information, a facet based model in which every facet adds an additional constraint to limit the result and where the results are combined using an AND function is most appropriate. Thus we have:

- *Filter variable:* in this role the variable V_{filt} together with a predicate P defines an element (V_{filt}, P) of a faceted filter of nominal labels, numeric range, concept score range, tags, time period, and buckets.

Based on the set of filter variables we thus define the active set of elements:

- *Active Set:* $M_{\text{active}} = \bigcap_k P^k_{\text{filt}}(V^k_{\text{filt}})$

The active set plays an important role in our approach as all subsequent pivoting steps take this set as starting point so that e.g. the weights of tags or the most relevant concepts are always determined in an active set specific manner.

To decide what elements can be row or column variables, we make two observations. First, there can be thousands of individual images and a magnitude more tags. Second, numeric values and thus also concepts can not be enumerated, they do induce an ordering though. Enumerating individual images or tags in both columns and rows would yield an explosion of cells. Therefore, we put such enumeration only as rows. This leaves the role for the column variable free and therefore we use variables put there to *sort* and possibly *weight* the data. Decomposition into the elements to use and the way of sorting defines which variables can be placed where. Location data cannot be meaningfully enumerated nor can they be used for sorting or weighting. So they can not be placed as row and neither as column variable. Nominal data can only be used as rows. Finally, numeric variables, temporal variables, and concepts can be used to sort and weight and hence are suited as column variable. A temporal variable as row variable is decomposed into natural time periods. Numeric variables, and concepts can be used as rows only after decomposing them into ranges. We do so by using 7-points summaries based on percentiles. How to create a 7-point summary depends on the shape of the distribution. For the highly skewed concept score distribution they are based on fixed percentiles denoted by p^{th}, where for numeric metadata exhibiting a more or less normal distribution we also provide the choice to having it based on the interquartile range $Q = 75^{\text{th}} - 25^{\text{th}}$ to define the normal range $[Q^-, Q^+]$ given by $[25^{\text{th}} - 1.5Q, 75^{\text{th}} + 1.5Q]$. Any data outside this range are considered statistical outliers. The two types of summaries are thus given as:

- *Fixed-7:* $(\min, 5^{\text{th}}, 25^{\text{th}}, \text{median}, 75^{\text{th}}, 95^{\text{th}}, \max)$
- *Derived-7:* $(\min, Q^-, 0.25^{\text{th}}, \text{median}, 75^{\text{th}}, Q^+, \max)$

So the row and column variables are defined as:

- *Row variable*: in this role the variable V_{row} defines the nominal values, tags, images, ranges of numeric or concept scores, or time periods to decompose M_{active} into the set $\mathcal{S}$ composed of the not necessarily disjunct subsets $\{S_j\}_{j=1..|\mathcal{S}|}$.
- *Column variable*: in this role the numeric, concept, or tag variable $V_{\text{sort-weight}}$ defines a weight w_{ij} corresponding to each $I_{ij} \in S_j$ to create an ordered sequence $\mathbf{S_j}$ with corresponding weight vector $\mathbf{w_j}$, whereas a temporal variable defines the ordering only.

Determining the weights is most difficult for tags. So let us consider that class of variables explicitly here. Let T_j denote the enumeration of all tags corresponding to the images in row S_j, let $t_{jk} \in T_j$ be the k_{th} unique tag and $f(t_{jk})$ the count of the tag in T_j. For a *frequency* weight we simply normalize $f(t_{jk})$ by $|T_j|$ the total number of tags. It is more interesting to highlight the set of tags in a cell which are typical for subset S_j with tags which are very common receiving less emphasis. The Okapi BM25 measure in information retrieval is designed for computing the relevance of keywords in text documents with respect to a query. It is directly applicable in our setting when we consider each tag as a query and consider the contribution of each tag t_{jk} within subset S_j. With $n(t_{jk})$ being the number of rows in which the tag occurs we can compute *BM25 weights* of tag importance. To that end let us first define the inverse term frequency measure IDF:

$$\text{IDF}(t_{jk}) = \log \frac{|\mathcal{S}| - n(t_{jk}) + 0.5}{n(t_{jk}) + 0.5} \tag{1}$$

Using this term we can define the weight as follows, with $\kappa = 1.6$ and $\beta = 0.75$ set to the recommended values:

$$w_{jk} = \text{IDF}(t_{jk}). \frac{f(t_{jk})(\kappa + 1)}{f(t_{jk}) + \kappa\left(1 - \beta + \beta \frac{|T_j|}{1/M \sum_j |T_j|}\right)} \tag{2}$$

We have finally arrived at assignment of variables to the value part of the pivot table. To allow to work with the diverse variables we are considering, we assign a variable to a column rather than to the whole table. In fact, we let users interactively define the variable to use in the value field and which variable to use as variable to sort and weight this particular column. So for each cell in the table we have a row variable defining which set of images to use for this particular cell, where each of them is giving a weight which also is providing the basis for sorting. When no column variable is used, we simply have a set of elements without a weighting function. All variables

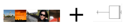
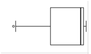
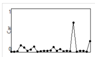
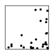
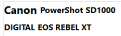
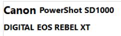
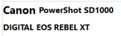
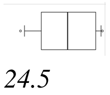
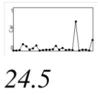
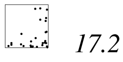
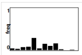
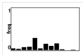
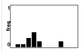
Variable					Visualization		
Type	**Filter**	**Row**	**Sort-Weight**	**Value**	**unsorted**	**sorted**	**weighted**
Images	Interactive selection	Individual images	x	List of images			
Concepts	Range selection	7-point summary (fixed)	Concept scores	Distribution			
Tags	Tag selection	Individual tags	BM25 / Frequency	Tag histogram	kaley rachel sidell gretchen sophie dog maggie dame notre de eiffel aigues beux bridge bus city cream gogh horse ice le mediterranean mortes riding salt sea sheep van walled zamor hall	kaley rachel sidell gretchen sophie dog maggie dame notre de eiffel aigues beux bridge bus city cream gogh horse ice le mediterranean mortes riding salt sea sheep van walled zamor hall	kaley rachel sidell gretchen sophie dog maggie dame notre de eiffel aigues beux bridge bus city cream gogh horse ice le mediterranean mortes riding salt sea sheep van walled zamor hall
Nominal	Label selection	Individual labels	x	Histogram			
Geo	Interactive selection	x	x	Set of coordinates			
Numeric	Range selection	7-point summary (fixed/derived)	Value	Sum, max, avg, distribution			
Buckets	Bucket selection	Individual buckets	x	Histogram			
Time	Period selection	Individual time periods	Temporal order	Time histogram			

TABLE I
Design rules for the different roles of variables and their visualization in multimedia pivot tables.

can be used for the value fields, the difference lies in how to aggregate the data and especially how to visualize the result.

- *Value variable*: in this role the variable V_{val} defines the set of attribute values $\mathbf{V_j}$ corresponding to the items in the ordered sequence $\mathbf{S_j}$ and the appropriate aggregation and visualization operators acting on $\mathbf{V_j}$ and weight vector $\mathbf{w_j}$.

Each of the cells in the multimedia pivot table now has an aggregated set of images or attribute values corresponding to the item subset S_j which depending on the underlying variables are sorted. Up to this point we did not consider the ordering of the rows of the table. This is based on a relevance function $R_\rho(V_{\mathrm{sort-weight}}, V_{\mathrm{val}})$ which is a function giving a value between -1 and +1 indicating the relevance of the items S_j in row j according to the two variables. Currently they are the Pearson correlation and if the sort-weight variable is empty it is simply any of the characteristics of the 7-point summary of V_{val} for the set S_j. The parameter ρ indicates the minimum number of items a row should have to be part of the table.

Based on the above we can conveniently describe a pivot table T in the following way:

$$T := ((V_{\mathrm{filt}}, P)^*, (V_{\mathrm{row}}, R_\rho), (V_{\mathrm{sort-weight}}, V_{\mathrm{val}})^*)$$

Where V denotes one of the variables, P is a predicate specification, and R is the relevance function. Note that the function R or the sort-weight variable can also be empty which gives a default sort order, e.g., by item ID or alphabetically. As an example, the table in figure 1 has the following specification:

$$\begin{aligned} T \;=\; & ((car, \mathrm{Between}(0.05, 1), (tags, \mathrm{Contains}(\text{“}car\text{”})) \\ & (ownername, \mathrm{Median}_5(indoor, -)), \\ & ((-, indoor), (outdoor, image), (-, road), \\ & (building, image), (-, tags), (-, datetaken), \\ & (-, geocoord), (datetaken, geocoord), \\ & (outdoor, road), (-, buckets)) \end{aligned}$$

D. Visualization

In a pivot table each cell corresponds to a set of items with the type of sort-weight and value variable determining the constraints on how to use this set. To allow for quick comparisons of the different subsets defined by the rows, small multiples are an appropriate mechanism to show patterns at a glance. For numeric variables, we can choose to show basic statistics like max, sum, or average. To get better insight in the values a more interesting choice is making use of a boxplot based on the 7-point summary. When a sort-weight variable is added, lineplots and scatterplots reveal trends and correlations, respectively. We observe that basic statistics reveal too little of the distribution of the concepts, while all the other visualizations are useful. To see the top ranked images according to the sort variable, the ranked list is simply shown with as many images as fit in the current column width. To get an understanding of the distribution of the weights of the images, an additional column showing the distribution can be added.

Visualizing the locations of items on a map reveals their spatial distribution. When combined with a temporal variable to sort the items, we can observe movement patterns. Finally mapping the weights to size we can also see how the importance of items is distributed. Temporal variables are shown as a time histogram with a natural binning in suitable intervals such as days or months. To visualize tags, text, and nominal variables, we use a simple tag cloud mechanism where important elements appear proportionally larger. As for both geo and tags the weights might vary considerably, and can be very different for each individual cell, we employ a 7-point summary to map the values to 7 different radii or font sizes.

Finally, to reveal the distribution of elements over the buckets they are visualized as a bar chart where the weights are accumulated per bucket. To reveal the distribution of number of elements per bucket the non-weighted distribution is visualized as a stacked bar chart in the horizontal direction. When weighed the vertical direction reveals the weight distribution normalized by the maximum value for the particular row.

The whole set of possible assignments of variables to the roles filter, row, column, or value and their associated visualizations are summarized in table I.

All columns perform *column aggregation* to have their totals over all elements visualized in the same manner as the individual cells to allow for assessments of an individual cell with respect to the whole. *Row aggregation* is performed to show the value and name of the maximal scoring concept over all concepts per row to define the dominant concept for each row-defined subset of the data.

When a relevance function is used for any of the columns, the background color of the cell shows the relevance to the user. We do so by mapping the relevance to a bi-polar color progression using luminance in the HSL color space, yellow for positive, and blue for negative values.

E. Interactive Analytics using Multimedia Pivot Tables.

User insight gain is facilitated through interactive browsing of the collection viewing a multitude of different multimedia pivot tables to find sets of interesting items in the collection, structure, patterns, and serendipitous findings. We now describe how we support the user with this process which is illustrated in figure 2.

For filtering it is important to find a natural way to specify the IDs of individual elements in M based on one of the variables. For nominal variables this is simply by filtering based on one or more specific labels. The same holds for keeping only the elements in a specific bucket. Both numeric metadata and concept scores can be filtered by using a range on the values. For concepts it should be noted that the cut-off point is often not clear. Hence, when using the ranking for filtering out items where the image doesn't contain the concept, thresholding the list should be done with care. We therefore support the selection by a simple interactive histogram based visualization. Tags or other textual descriptions can be selected through regular expressions. Specifying geolocation by ranges is cumbersome so we provide a map visualization in which users can select elements by dragging a rectangular region. Selecting images through other means than concepts or attributes can only be done via visual inspection. We do so by letting the user interactively select them in any of the visualizations. The user can then put the selection in a bucket for filtering, pivoting with buckets as variable, or any other future reference.

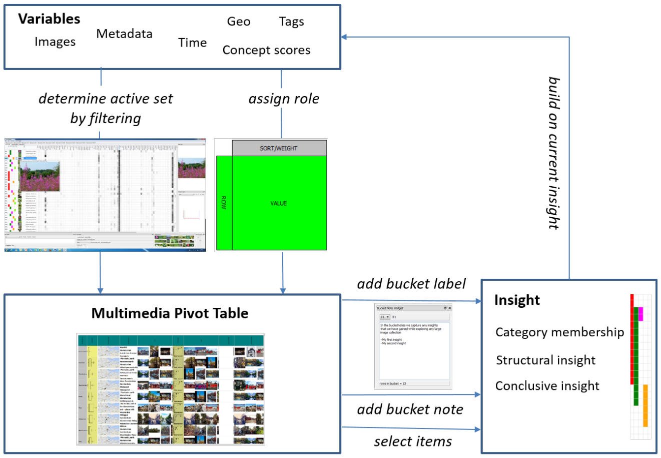


Fig. 2. The multimedia analytics process.

A visual widget supports the assignment of variables to the other roles a variable can have (row, sort-weight, value). The widget employs the design rules as described in table I to assure that users can only assign variables to valid roles. Proceeding in this manner, the users can interactively create and browse different multimedia pivot tables until some insights in the collection have been found.

The first insights to expect are structural. Thus, the user starts labeling one or more buckets so that they become containers with a semantic meaning to which elements can be added. Through mouse clicks the user can select sets of images by taking individual rows in the pivot tables or any of the other visualizations (like the standard table based visualization, the grid based visualization, or the image scatterplot). The user can add the selected elements to one or more of the buckets with simple keyboard shortcuts, or remove the selected elements from the bucket if their bucket membership needs reconsideration. The user can also negate the membership of all members in a specific bucket. As buckets can also be used in the faceted filter which is based on the boolean AND, the full richness of set operations are available to the user.

Having captured the first structural insights through the buckets having received their first meaning and members, the user can build upon them by giving the bucket variables pivot roles to reveal differences and similarities among the different groups in the collection. As soon as the user finds conclusive insights they can be added in the form of bucket notes attached to the buckets after which the user can continue building upon what was found. In the different interaction rounds the user builds upon structural and conclusive insights. As this might be a lenghty process note that all the essentials of the state are captured by the current insights B^t and pivot table T. The logging function of the system stores these insights and the current pivot table and the whole sequence can be repeated to reveal how and at what times insights have been obtained.

With the insights and the process of obtaining them available, true understanding of the collection might be in sight.

IV. Implementation

Our implementation of the multimedia pivot table paradigm is based on our offline C++ library to compute concepts based on a deep learning pipeline, wheras Qt is used for the online system. Qt's intelligent scrolling mechanisms automatically adjust the number of rows displayed to allow working with large numbers of rows (we tested up to one million) and hundreds of columns and which only requires visualizations for visible cells. For the pivot tables in addition the use of filter variables and the ρ parameter give good mechanisms to only show the most relevant items. Most expensive operation is the use of tags as their number is an order of magnitude larger than the number of images. An essential idea behind the tables is that tag distribution cannot be pre-computed as they are re-computed whenever the active set is changed. To assure responsivess of the interface the implementation is multi-threaded so that the interface doesn't block when waiting for an operation to finish.

V. Use Case Scenarios and Evaluation

To evaluate the multimedia analytics capability of pivot tables, we have conducted a number of user studies following the open-ended insight-focused protocol of North [31][32]. The first set of user studies was conducted with novice users to get feedback on our initial design while the second set of user studies was with expert users being the target group of the tool. We will now first illustrate two use-case scenarios of typical situations where multimedia pivot tables could show their use.

A. Use case scenarios

Consider a social scientist analyzing the photo sharing site Flickr aiming to understand what makes photos popular. He starts off by using a 7-point summary to decompose the set into rows based on the *number of likes* a picture has. Looking at the aggregated values for the *owner* variable reveals that the positive outliers are mostly from a small set of photographers. The scientist filters them out for later study and continues with the rest. The updated pivot table shows that for the most liked pictures many *tags* like “awesome” and “amazing” are used, but also tags related to landscapes. So he adds the *landscape* concept and also adds a column with *images* sorted by their concept scores for *landscape*. Many top images are from mountains so he adds the *mountain* concept and at the same time adds the variable *geocoordinates* to show their spatial distribution which is concentrated in the Rocky Mountains and the Alps. As he also expects time to make a difference, he uses *dateupload* to decompose the set in rows for individual months. Looking at the sorted images he now identifies a number of interesting geotemporal subsets to study further.

As a second use case we consider an intellligence agent analyzing the mobile phone data from a terrorist suspect. As a starting point she creates three buckets containing images with high scores for the concepts *weapon, building, bridge,* and *train* respectively. Setting these three *buckets* as row variable she adds *time* as value variable which gives a clear indication that there are two distinct periods in which bridges have been studied. Using *time* as row variable and *geocoordinates* as value variable also reveals that they are all from three closely related geolocations. Inspecting the corresponding images indeed show three different bridges. One of the groups also shows a strong presence of the *train* detector so the investigator hypothesizes that there might be a plan for an attack on a train while crossing a bridge and adds this as one of the possible scenarios to investigate.

B. Datasets

For the experiments we consider three different datasets.

The *Flickr dataset* is a set of 17K images publicly available from Flickr using a set of 20 query terms for specific types of objects and scenes. The gathering was done such that all resulting images have geocoordinates. In addition to the images, we collected metadata such as number of views, title, owner name, tags, camera used, and camera parameters. We apply a set of 150 concept detectors using the techniques in [37] which are trained on data outside of the current dataset. This set of detectors is a good compromise between providing sufficient accuracy while providing sufficient choice to select a small set of detectors relevant to the given task and query.

The *IAmsterdam dataset* and *Fastfood dataset* are both crawled from Instagram and focus on brand presence in social media. *IAmsterdam* has 7K items and focuses on the IAmsterdam letters used to promote Amsterdam[2]. For this dataset a small set of domain-dependent detectors is trained, namely the following concepts: IAmsterdam, Rijksmuseum, bike, and canal. The *Fastfood* dataset is composed of 53K items obtained by querying for the hastags related to the most popular fast food brands. In addition to the Instagram metadata attributes, they are described with the 1200 learned adjective-noun pairs from [2].

C. User study with novice users

Participants of the first experiment were a group of 34 master students in information science and artificial intelligence as well as 7 PhD students in various technical disciplines working on the *Flickr dataset*. Each student worked between 2 and 3 hours in total on the task. Part of this time was devoted to an intro by the instructor and some free exploration of the tool before notes had to be written down by the participants.

Following the guidelines in [31], the participants were not given a particular task, they were rather freely exploring the dataset for insights of interest. The participants were given a form in which they indicated what categories/buckets they used at the start and which new categories they found. After the experiments they could comment on the weak and strong points of the tool and make suggestions for further improvement.

Each of the participants came to between two and eight observations of interest. Many of them were related to the

[2] iamsterdam.com

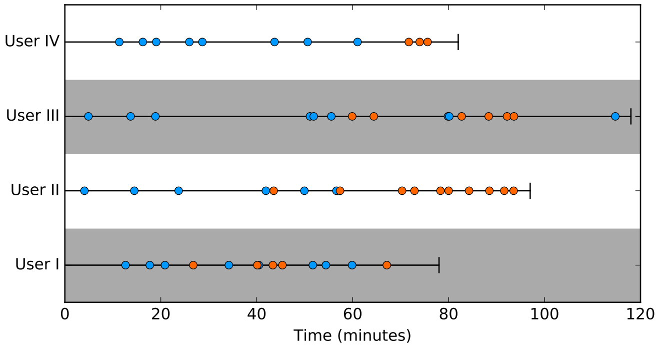


Fig. 3. Insight gains by experts over time. Blue dots denote structural insights, orange dots denote conclusive insights. Users with white background are multimedia analysis experts, users with gray background are marketing professionals.

performance of detectors, where they worked and where they didn't. We analyzed all observations the participants wrote down and the different ways in which they reached these insights. Within the relatively short time given to work with the tool, most students were finding some simpler forms of insight. A few students were able to indeed find observations requiring complex relations between different columns and rows. We have used all comments to improve the initial design (see [45] for some of the changes), leading to the system used in the advanced user study.

D. User study with experts

For the advanced user study, we have engaged 4 experts: 2 marketing professionals (users I and III) and 2 postdoctoral researchers (users II and IV) working on multimedia analysis. Users I and II have explored the *IAmsterdam dataset* while users III and IV explored the *Fastfood dataset*.

Each user was asked to explore the multimedia dataset based on his expertise and report the attained insights via the bucket notes. The users were not limited with respect to time spent in the evaluation session. We have recorded the two types of insight: *structural insight* and *conclusive insight* as defined in Section III-B.

The insight gains over time are depicted in figure 3 whereas details on the insights themselves are provided via our website. The graph and insights indicate that pivot tables succeed in comparison with using a sequential, non-summarization-based exploration approach. Indeed, even if a user would be able to process 1 item per second, he would be able to see only a fraction of the explored dataset in an hour. Using pivot tables, the users were able to attain both structural and conclusive insights within the first hour of the analysis. The fast insight gain also confirms the hypothesis about spreadsheets being a familiar analytic metaphor: the learning curve is not crippling.

In multimedia analytics, it is imperative that a successful approach allows for insights based on high-level semantic information. Most insights gained by the users in the evaluation are indeed semantic, such as "I was looking for images of food and it surprised me that most of the images were of healthy food" or "canal images are the most popular (liked) Amsterdam images, with only 9/34 in the top featuring the iAmsterdam letterset." Pivot tables thus succeed in providing meaningful semantic summarization of the collection.

After the analysis, each user has been asked about the main perceived strengths and weaknesses. The users were positive about the approach, stating that it is a powerful visualization able to convey large amounts of data in an understandable way. The users also commended the flexibility in selection of the dimension along which the exploration is guided, as well as the ability of multimedia pivot tables to combine a large number of useful visualizations in a seamless manner. The weaknesses revolved chiefly about adjustments to the interface increasing user friendliness and the imperfections in data annotation. The suggestions about user friendliness included for example better labeling of some of the plots (chiefly histograms) and the ability to search attribute names. Regarding data annotations, the users stated that they would welcome more content annotations of the data, so that they could explore the data further. The main perceived weaknesses of the system were insufficient support of data selection directly in the cells of the pivot table; inability to select valuable discriminative attributes automatically; and responsiveness in particular for tags in the case of the bigger *Fastfood* dataset. These aspects serve as an interesting direction for further research. Overall, pivot tables were deemed a powerful approach for multimedia analytics by the users: all of them commended the analytic capability and most of the perceived weaknesses were suggestions for improvement, rather than aspects impeding the analytic process.

VI. Discussion

Multimedia pivot tables provide a powerful mechanism to obtain structural and conclusive insight in image collections. Where search based approaches only show one ranking, our filter/search interface provides many at the same time. Yet only with pivot tables we get group based statistics and aggregated results over those. The current system does have its limitations though, and most of these would require additional analysis and machine learning methods. We have a strong focus on the visual content and row aggregations are exclusively based on the visual concepts. When the tags would e.g. be analyzed using topic models we could also aggregate over multiple tag sets and even perform multimodal aggregations over both visual content and text. Furthermore, we use large concept sets (currently ranging to more than 10,000). When moving to such large sets users should get support in selecting the right concepts (or textual topics) for the active datasets and maybe even for each row. In addition, we emphasize group membership to compare sets of images, similarity based visualizations would bring out additional patterns which might be difficult to find using pivot tables. Finally, selecting elements for the buckets is a manual and time-consuming process. Active buckets [8] as used in Mediatable which give data-driven suggestions for bucket membership could substantially improve this process, in particular if it would go beyond the visual content alone.

VII. Conclusion

We have proposed a new multimedia analytics solution which takes the summarization power of pivot tables so common in spreadsheets as basis and brings them into the realm of multimedia and its associated analytic processes.

The characteristics of different types of variables yield the basis for what pivot roles variables can take, where we have identified the additional role of sort-weight and added a relevance function to rank the rows in the pivot table. These assignments of variables give us the means to interactively perform multimodal summarization where specific visualizations support the interpretations of individual cells in the table. The variable assignment rules and visualizations of cells are depicted in table I. To make multimedia pivot tables support the structuring part of the analytics process, we have explicitly defined and operationalized structural and conclusive insight in terms of category labels, item membership, and conclusions on those. By embedding them in the multimedia analytics process depicted in figure 2 we get a highly interactive process to browse image collections in search of patterns and subsequent insights which addresses to varying extent all dimensions of insight identified in [31]:

- *Complex:* Multimedia pivot tables allow any combination of the variables in the different roles, only constraint by the assignment rules of table I, hence they are truly multimodal and provide summarization to aggregate underlying data and to see relations.
- *Deep:* The analytics process (figure 2) is designed to built upon the currently gathered category membership, structural, and conclusive insights.
- *Qualitative:* Category labels are qualitative by definition and the categories can be dynamically annotated with conclusive insights.
- *Unexpected:* The highly flexible variable assignments lead to easy generation of user defined summaries integrating various dimensions and allowing to see relations and patterns not revealed by simple query based results.
- *Relevant:* The interactive analytics process and insight gathering is fully in the hand of the expert, directly reflecting domain dependent terms and hence expected to lead to direct domain impact.

The user experiments, especially with the expert users, reveal that indeed users are capable of getting both structural and conclusive insight in the collection by using multimedia pivot tables. Our expert users did so by using the different modalities and their insights were at a highly semantic i.e. qualitative and domain relevant level. The insight gain plots (figure 3) show that insights found were alternating between structural and conclusive. All of these are indications that the proposed multimedia analytics solution contributes to all the dimensions of insight gain. Clearly the experiments were based on short sessions so the insights are still simple and conclusions preliminary. The tool is currently deployed in a number of projects and through extended use we hope to determine how well the methods perform when used over prolonged periods of time with specific datasets and specific domain experts.

Full multimedia analytics will remain a challenge for many years to come, with many interesting research challenges [47][19] and we are currently scratching the surface of its full potential. Our solution has made a contribution by not addressing one of the insight characteristics, but making a step towards the integral support of different ones.

Acknowledgements

We would like to thank all the participants in the user experiments. In particular Olivier Ponti, Masoud Mazloom, Stevan Rudinac, and Bob Rietveld.

References

[1] W. Aigner, S. Miksch, H. Schumann, and C. Tominski. *Visualization of Time-Oriented Data*. Springer Publishing Company, Incorporated, 2011.

[2] D. Borth, R. Ji, T. Chen, T. Breuel, and S.-F. Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *ACM Multimedia*, 2013.

[3] R. Burtner, S. Bohn, and D. Payne. Interactive visual comparison of multimedia data through type-specific views. In *SPIE, Visualization and data analysis*, 2013.

[4] Q. Chen, Y. Chen, D. Liu, C. Shi, and Y. Wu. Peakvizor: Visual analytics of peaks in video clickstreams from massive open online courses. *IEEE Transactions on Visualization and Graphics*. 2015.

[5] E. Chi, J. Riedl, P. Barry, and J. Konstan. Principles for information visualization spreadsheets. *Computer Graphics and Applications*, July/August 1998.

[6] N. Chinchor, J. Thomas, P. C. Wong, M. Christel, and W. Ribarsky. Multimedia analysis + visual analytics = multimedia analytics. *Computer Graphics and Applications, IEEE*, 30(5):52–60, 2010.

[7] M. Christel. *Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation*. Morgan and Claypool Publishers, 2009.

[8] O. de Rooij and M. Worring. Active bucket categorization for high recall video retrieval. *IEEE Transactions on Multimedia*, 15(4), 2013.

[9] O. de Rooij, M. Worring, and J. J. van Wijk. Mediatable: Interactive categorization of multimedia collections. *IEEE Computer Graphics and Applications*, 30(5):42–51, 2010.

[10] J. Deng, A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *Proceedings of the 12th European Conference of Computer Vision (ECCV)*, 2010.

[11] N. Elmqvist and J.-D. Fekete. Hierarchical aggregation for information visualization: Overview, techniques, and design guidelines. *IEEE Transactions on Visualization and Computer Graphics*, 16(3):439–454, 2010.

[12] D. Fisher, S. Drucker, R. Fernandez, and S. Ruble. Visualizations everywhere: A multiplatform infrastructure for linked visualizations. *IEEE Transactions on Visualization and Computer Graphics*, 16(6), 2010.

[13] A. Girgensohn, F. Shipman, T. Turner, and L. Wilcox. Flexible access to photo libraries via time, place, tags, and visual features. In *Proceedings of JCDL*, 2010.

[14] S. Gratzl, A. Lex, N. Gehlenborg, H. Pfister, and M. Streit. Lineup: Visual analysis of multi-attribute rankings. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2277–2286, 2013.

[15] J. Gray, S. Chaudhuri, A. Bosworth, A. Layman, D. Reichart, and M. Venkatrao. Data cube: a relational aggregation operator generalizing group-by, cross-tab, and sub-totals. In *Data Mining and Knowledge Discovery*, 1997.

[16] B. Höferlin, R. Netzel, M. Höferlin, D. Weiskopf, and G. Heidemann. Inter-active learning of ad-hoc classifiers for video visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST), 2012*, pages 23–32, 2012.

[17] B. Jelen and M. Alexander. *Pivot Table Data Chrunching*. Que Publishing, 2005.

[18] P. Joia, F. Paulovich, J. C. D. Coimbra, and L. Nonato. Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2011.

[19] B. Jónsson, M. Worring, J. Zahálka, S. Rudinac, and L. Amsaleg. Ten research questions for multimedia analytics. In *Multimedia Modeling*, 2016.

[20] S. Kandel, E. Abelson, H. Garcia-Molina, A. Paepcke, and M. Theobald. Photospread: A spreadsheet for managing photos. In *Proceedings of SIGCHI*, 2008.

[21] D. Keim, G. Andrienko, J.-D. Fekete, C. Gorg, J. Kohlhammer, and G. Melancon. Visual analytics : Definition, process, and challenges. In *Information Visualization, LNCS 4960*, 2008.

[22] D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann, editors. *Mastering The Information Age - Solving Problems with Visual Analytics*. Eurographics, Nov. 2010.

[23] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan. Large-scale video summarization using web-image priors. In *Computer Vision and Pattern Recognition*, 2013.

[24] G. Kim, L. Sigal, and E. Xing. Joint summarization of large-scale collections of web images and videos for storyline reconstruction. In *Computer Vision and Pattern Recognition*, 2014.

[25] K. Kucher and A. Kerren. Text visualization browser: A visual survey of text visualization techniques. In *Proceedings of INFOVIS*, 2014.

[26] Y. LeCunn, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521:436–444, 2015.

[27] H. Luo, J. Fan, S. Satoh, J. Yang, and W. Ribarsky. Integrating multi-modal content analysis and hyperbolic visualization for large-scale news video retrieval and exploration. *Signal Processing: Image Communication*, 23(7), 2008.

[28] S. MacNeil and N. Elmqvist. Visualization mosaics for multivariate visual exploration. *Computer Graphics Forum*, 2013.

[29] T. Mei, Y. Rui, S. Li, and Q. Tian. Multimedia search reranking: A literature survey. *ACM Computing Surveys*, 46(3), 2014.

[30] G. P. Nguyen and M. Worring. Interactive access to large image collections using similarity-based visualization. *Journal of Visual Languages and Computing*, 19(2):203–224, 2008.

[31] C. North. Toward measuring visualization insight. *IEEE Comput. Graph. Appl.*, 26(3):6–9, 2006.

[32] C. North, P. Saraiya, and K. Duca. A comparison of benchmark task and insight evaluation methods for information visualization. *Information Visualization*, 10(3):162–181, 2011.

[33] D.-S. Ryu, W.-K. Chung, and H.-G. Cho. Photoland: a new image layout system using spatio-temporal information in digital photos. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, SAC '10, pages 1884–1891, 2010.

[34] K. Schoeffmann, D. Ahlström, and L. Böszörmenyi. 3D storyboards for interactive visual search. In *ICME*, 2012.

[35] K. Schoeffmann, M. A. Hudelist, and J. Huber. Video interaction tools: a survey of recent work. *ACM Computing Surveys*, 2015.

[36] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, 2000.

[37] C. G. M. Snoek and et.al. Mediamill at TRECVID 2013: Searching concepts, objects, instances and events in video. In *Proceedings of TRECVID Workshop, Gaithersburg, USA*, 2013.

[38] C. Stolte, D. Tang, and P. Hanrahan. Polaris: a system for query, analysis, and visualization of multidimensional relations databases. *IEEE Transactions on Visualization and Computer Graphics*, 8(1), 2002.

[39] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Computer Society, 2005.

[40] G. Tómasson, H. Sigurthórsson, B. Jónsson, and L. Amsaleg. Photocube: effective and efficient multi-dimensional browsing of personal photo collections. In *Proceedings of the ICMR*, 2011.

[41] B. Truong and S. Venkatesh. Video abstraction: a systematic review and classification. *ACM Transactions on Multimedia Computing Communications and Applications*, 3(1), 2007.

[42] P. van der Corput and J. van Wijk. ICLIC: Interactive categorization of large image collections. In *IEEE Pacific Visualization Symposium*, 2016.

[43] C. Wang, J. Reese, H. Zhang, J. Tao, Y. Gu, J. Ma, and R. Nemiroff. Similarity-based visualization of large image collections. *Information Visualization*, 6, 2013.

[44] C. Ware. *Visual Thinking for Design*. Morgan Kaufmann, 2008.

[45] M. Worring and D. Koelma. Insight in image collections by multimedia pivot tables. In *ACM International Conference on Multimedia Retrieval*, 2015.

[46] J. Yang, J. Fan, D. Hubball, Y. Gao, H. Luo, W. Ribarsky, and W. M. Semantic image browser: Bridging information visualization with automated intelligent image analysis. In *IEEE Symposium on Visual Analytics Science and Technology*, 2006.

[47] J. Zahálka and M. Worring. Towards interactive, intelligent, and integrated multimedia analytics. In *IEEE Conference on Visual Analytics Science and Technology*, 2014.

[48] E. Zavesky, S.-F. Chang, and C.-C. Yang. Visual islands: Intuitive browsing of visual search results. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*, pages 617–626, 2008.

Marcel Worring received the M.Sc. degree (Hons.) in computer science from VU University Amsterdam, Amsterdam, The Netherlands, in 1988, and the Ph.D. degree in computer science from the University of Amsterdam, Amsterdam, The Netherlands, in 1993. He is currently an Associate Professor within the Informatics Institute, University of Amsterdam, of which he is the director, and a Full Professor with the Amsterdam Business School, University of Amsterdam. He has (co-)authored over 170 scientific papers covering a broad range of topics. His research interest is the integration of multimedia analysis, multimedia mining, information visualization, and multimedia interaction into a coherent framework yielding more than its constituent components. Prof. Worring was an Associate Editor of the IEEE TRANSACTIONS ON MULTIMEDIA and is currently an Associate Editor of the ACM Transactions on Multimedia. He was Program Chair for ICMR 2013 and ACM Multimedia 2013, and is general chair for the ACM Multimedia 2016.

Dr. Dennis Koelma is senior scientific programmer at the UvA. He received the M.Sc. and Ph.D. degrees in computer science from the University of Amsterdam in 1989 and 1996, respectively. His research interests include image and video processing, software architectures, parallel programming, databases, graphical user interfaces, and visual information systems. He is the lead designer and developer of Impala: a software architecture for accessing the content of digital images and video. The software serves as a platform for consolidating software resulting from the research group. It has been licensed by the UvA spin-off Euvision where he had a part-time affiliation until it was acquired by Qualcomm.

Jan Zahálka received the M.S. degree in artificial intelligence from the Czech Technical University, Prague, Czech Republic, in 2013, and is currently working toward the Ph.D. degree at the University of Amsterdam, Amsterdam, The Netherlands. His research interests include multimedia analytics, namely integrating heterogeneous information contained in different modalities using interactive machine learning and visual analytics techniques.