Keep Moving!
Revisiting Thumbnails for Mobile Video Retrieval

Wolfgang Hürst
Utrecht University
PO Box 80.089
3508 TB Utrecht
The Netherlands

Cees G. M. Snoek
University of Amsterdam
Science Park 107
1098 XG Amsterdam
The Netherlands

Willem-Jan Spoel
Utrecht University
PO Box 80.089
3508 TB Utrecht
The Netherlands

Mate Tomin
Utrecht University
PO Box 80.089
3508 TB Utrecht
The Netherlands

huerst@cs.uu.nl, cgmsnoek@uva.nl, cwjispoe@students.cs.uu.nl, mtomin@students.cs.uu.nl

ABSTRACT

Motivated by the increasing popularity of video on handheld devices and the resulting importance for effective video retrieval, this paper revisits the relevance of thumbnails in a mobile video retrieval setting. Our study indicates that users are quite able to handle and assess small thumbnails on a mobile's screen – especially with moving images – suggesting promising avenues for future research in design of mobile video retrieval interfaces.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**]: User Interfaces – *Evaluation/methodology, graphical user interfaces (GUI), screen design, style guides, user centered design*

General Terms

Design, Experimentation, Human Factors.

Keywords

Mobile video, video retrieval interfaces, visual assessment tasks.

1. INTRODUCTION

Despite the increasing importance of effective video searching on mobiles, surprisingly few retrieval interfaces have been optimized for interaction on handheld devices [1, 3]. The question how advanced video browsers should be adapted to the constraints imposed by a mobile's small screen while maintaining usability remains unanswered. In this paper, we take a closer look at the basic building block of such interfaces: the thumbnail, i.e. reduced-size versions of a single static image (subsequently called ***static thumbnails***) or a set of moving images (subsequently called ***dynamic thumbnails***) that are extracted from a short piece of video – usually a camera shot assumed to be representative for its content. Contrary to widespread believe that a mobile's small screen size would not allow displaying such without loss of recognition, Torralba *et al*. [5] revealed that humans are able to outperform computer vision algorithms in an image recognition task on the desktop, even at strongly reduced versions of the original images. Motivated by these perceptual findings, we present experiments where subjects had to perform verification tasks based on a single thumbnail extracted from a video – a common task in video retrieval, for example when assessing the relevance of a search result. Our goal is to evaluate the importance of thumbnail sizes and the influence of static versus dynamic thumbnails for human recognition performance.

MM'10, October 25–29, 2010, Firenze, Italy.

2. HUMAN-CENTERED EXPERIMENTS

All experiments have been done with a *Motorola Droid* phone running Android OS version 2.0 (cf. Fig. 1). It features a touch screen with a relatively large resolution of 854x480 pixels. Although we expect most phones to increase in screen resolution in the future, this is clearly above the current state-of-the-art and at the upper end – even for smart phones. Therefore, we decided to implement and run all experiments in compatibility mode with older Android OS versions, resulting in a screen resolution of 569x320 pixels that was used in all tests.

2.1 Two User Study Experiments

We set up two experiments in which the participants had to assess the relevance of a typical video retrieval result based on a single thumbnail at various sizes. Inspired by the work of Torralba *et al*., who found that images of 32x32 pixels were often sufficient to recognize the content of images on the desktop [5], we set the minimum thumbnail width at 30 pixels and incremented it successively with 10 pixels until a width of 120 pixels is achieved – which is a typical size of a thumbnail in traditional video retrieval interfaces, as used in the TRECVID video retrieval benchmark [4]. The height of the images is adapted according to the video's aspect ratio. However, we realized that human recognition performance even at 30 pixels is extremely high when the device is hold unnaturally close to ones face. Therefore, participants were asked to "hold the device in a natural and comfortable way", for example by resting their arms on a table (cf. Fig. 2). A neutral observer reminded them of this guideline when an awkward position was recognized during the evaluations.

Overall, 24 users (22 male, 2 female, ages 1 from 15-20, 17 from 21-30, 3 from 31-40, and 3 from 41-50) participated in both experiments – subsequently referred to as *experiment A* and *B*. Half of the subjects started with experiment A followed by B, the other with B followed by A. Experiments have been done in a quiet place with no distractions and subjects sitting comfortably on a chair. Videos and thumbnails were taken from the TRECVID benchmark [4], and realistic questions were selected from [2]. Some questions needed to be adapted in order to fit to a "yes/no" answer scheme (which was chosen to focus on the independent

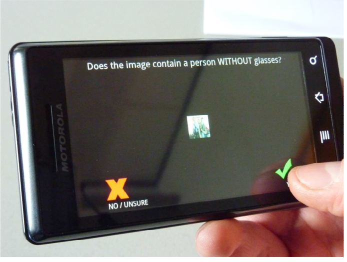
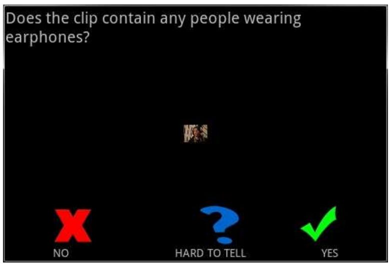


Figure 1. Mobile phone used in the studies (here: experiment B with random thumbnail sizes).

Figure 2. User study participants while assessing the relevance of video retrieval results on the mobile phone.



Does the clip contain any people wearing earphones?

Figure 3. Interface in experiment A, with different thumbnail sizes (here: smallest, i.e. 30 pixels width).

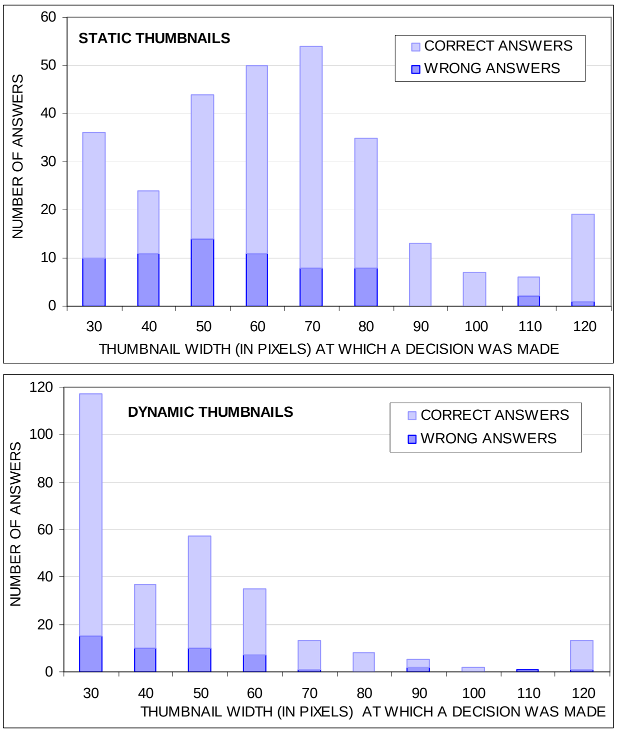


Figure 4. Results for experiment A: Thumbnail Size Preference. Human performance when assessing the relevance of video retrieval results on a mobile phone, using increasing thumbnail sizes for static (top) and dynamic thumbnails (bottom). Note the ease with which users are able to correctly classify small dynamic thumbnails.

variables thumbnail size and type; cf. below) but were similar in spirit to the ones used in the literature. Questions were chosen randomly, but under consideration of covering different retrieval tasks – in particular: object and subject verification (e.g. “Does the clip contain any police car?”) versus scene and event verification (e.g. “Does the clip contain any moving black car?”).

Experiment A: Thumbnail Size Preference. The major goal of experiment A was to evaluate at what thumbnail sizes people feel most comfortable and confident when making their decision. For this, the participants had to answer 24 questions (12 with static thumbnails, 12 with dynamic). Half of the subjects started with static thumbnails, half with dynamic ones. Dynamic thumbnails were played in an endless loop. Users first saw the smallest thumbnail size (30 pixels width) and could give their answer (by hitting “yes” or “no”) or hit another button (labeled with “?” and “hard to tell”) to enlarge the thumbnail by 10 pixels up to the maximum width of 120 pixels (cf. Fig. 3). They were asked to make a choice at the smallest possible thumbnail size at which they felt confident to make a correct decision. To motivate users to decide at smaller thumbnail sizes, we enforced a slight delay after they pushed the enlargement button before allowing them to further increase thumbnail size. To eliminate any interference with the decision process the background was set to black.

Experiment B: Thumbnail Size Influence. The purpose of experiment B was to evaluate human performance at varying sizes of static and dynamic thumbnails. For this, each participant had to answer another 24 questions (again 12 based on static thumbnails and 12 on dynamic; half of the users starting with static ones, half with dynamic ones). Questions and data were different than in experiment A but created in a similar way. Thumbnail sizes were again restricted to widths of 30, 40, 50, 120 pixels. However, this time they were presented in random order and there was no possibility to modify their sizes, but users had to make a decision based on the given size and type. Hence, the interface had only two buttons (“yes” and “no/unsure”, cf. Fig. 1).

2.2 Results

Experiment A: Thumbnail Size Preference. We plot the results of experiment A in Figure 4. Overall, we see a high number of decisions at relatively small sizes for both static and dynamic thumbnails. For static thumbnails, the majority of questions were answered for sizes smaller than 90 pixels, for dynamic ones sizes are typically below 70 pixels. The average size used for the final judgment when assessing static thumbnails was 64.7 pixels. Correct answers had an average size of 67.2 pixels, and wrong ones had an average of 56 pixels. Since users were able to increase the size, we can assume that they felt confident about giving a correct answer even when making a wrong decision. For dynamic thumbnails, average sizes are much lower. Moreover, there is hardly any difference between wrong and correct answers: average values are 48.0 pixels overall, 48.0 pixels for correct and 47.9 pixels for wrong answers.

We also observe that participants performed much better for dynamic thumbnails – with the total number of 65 wrong answers for static versus 47 for dynamic thumbnails. For static ones, people hardly made any mistake for sizes larger than 80 pixels. For dynamic ones, almost all answers for sizes larger than 60 pixels are correct. Despite the larger amount of mistakes at lower pixel sizes, the results reveal a relatively high number of correct

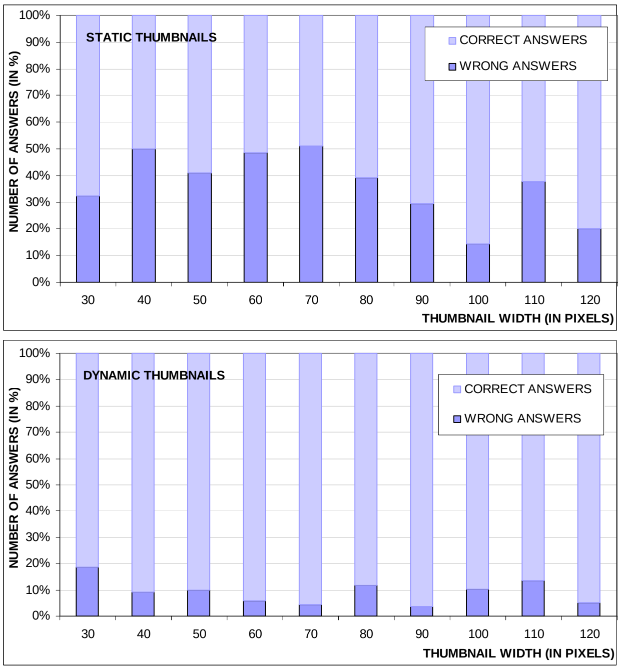


Figure 5. Results for experiment B: Thumbnail Size Influence. Human performance when assessing the relevance of video retrieval results on a mobile phone, using random thumbnail sizes for static (top) and dynamic thumbnails (bottom). Note the overall decrease in number of errors when assessing dynamic thumbnails.

answers at low sizes. Particular noteworthy is the high number of correct decisions with dynamic thumbnails, were 41% of all decisions have been made at the smallest thumbnail width of 30 pixels. What is more, 87% of those decisions have been correct. This result is especially surprising with respect to Torralba *et al*.'s findings for 32x32 sized images [5], because thumbnails extracted from videos are often of lower quality than individually created images and have a smaller height than width because of the video's aspect ratio.

Experiment B: Thumbnail Size Influence. Figure 5 illustrates the outcome of experiment B where thumbnail sizes were used in random order. Results are plotted by the accumulated total number of correct versus wrong answers for each assessed thumbnail size and type (static and dynamic).

Decisions made based on dynamic thumbnails clearly outperform the ones made with static ones of similar sizes. In addition, for dynamic thumbnails human performance does not decrease for smaller sizes. The results of experiment A show that the participants made almost no mistakes for sizes of 90-120 pixels for static thumbnails and 70-120 pixels for dynamic ones (cf. Fig. 4). Hence, we can assume that the amount of errors made at these sizes is a good indication for general human performance on the given data set. For static thumbnails, 25% of the decisions made at sizes 90-120 have been wrong. For dynamic ones, 8.5% of the decisions made at sizes 70-120 have been wrong (8.4% for sizes 90-120). Comparing these values with the errors made at lower thumbnail sizes (where we can assume that the size has an influence on human performance, cf. exp. A) shows a larger increase in errors for static thumbnails from 25.0% to 44.2% for sizes 30-80. In contrast to this, we see almost no difference for dynamic ones where error increases only from 8.5% to 10.5% for sizes 30-60 (and from 8.4% to 9.9% for sizes 30-80).

Hence, these results confirm our findings from experiment A that sizes for static thumbnails should be at least 90 pixels or higher for a good recognition performance on a mobile. In addition, we observe again a very good performance at much smaller sizes if dynamic thumbnails are used, with indications for an optimum thumbnail size being at least 70 pixels, but a surprisingly high performance already at thumbnail widths as low as 30 pixels.

2.3 Task-Dependent Performance

Our experiments reveal an obvious advantage of dynamic over static thumbnails because they enabled participants to achieve a better verification performance at lower thumbnail sizes. In order to further investigate these observations, we evaluated the results according to different video retrieval tasks, i.e. verification of objects/subjects versus scenes/events (cf. the description of the data set in 2.1). Intuitively, we would assume that dynamic thumbnails perform better on scene/event verification tasks because they preserve the dynamic nature of the respective information. Static ones might be better for object/subject verification because dynamic thumbnails can also include several frames where the object/subject is not clearly visible and thus introduce noise and create distraction.

Figure 6 illustrates the results from **experiment A** (cf. Fig. 4) split into object/subject tasks (left) versus scene/event tasks (right) for static (top row) and dynamic thumbnails (bottom row). Each of the four diagrams is based on 144 samples. Table 1 summarizes the average thumbnail sizes at which a decision was made. The data confirms the general trends identified before but also reveals important differences between the two task types. For example, Table 1 confirms the trend that people prefer larger sizes for static thumbnails compared to dynamic ones for both tasks. There is no notable difference in thumbnail sizes between correct and wrong decisions except for scene/event tasks with static thumbnails where many decisions made at smaller thumbnail sizes have been wrong (average thumbnail size 71 for correct versus 50 for wrong decisions) thus confirming our intuitive assumption that dynamic thumbnails are better for these kind of tasks. However, decisions made at static thumbnail sizes larger than 90 pixels have mostly been correct (cf. Fig. 6) indicating that even for scene/event tasks humans are able to make reliable decisions based on static thumbnails if they are large enough. For dynamic thumbnails, almost no mistakes were made for sizes larger than 70 pixels thus confirming the previously identified thresholds for all four thumbnail/task combinations.

Considering the absolute number of mistakes, participants made far less errors with dynamic thumbnails independent of task and thumbnail type. Comparing the number of mistakes made with dynamic thumbnails for object/subject versus scene/event tasks (26 vs. 14 mistakes) confirms our previously mentioned intuition that dynamic thumbnails could introduce noise that complicates the decision process. However, the assumption that static thumbnails might therefore be better for object/scene tasks was not confirmed since the number of errors on the comparable data set was much higher (39 vs. 26). Although the difference in performance between object/subject tasks and scene/event tasks is

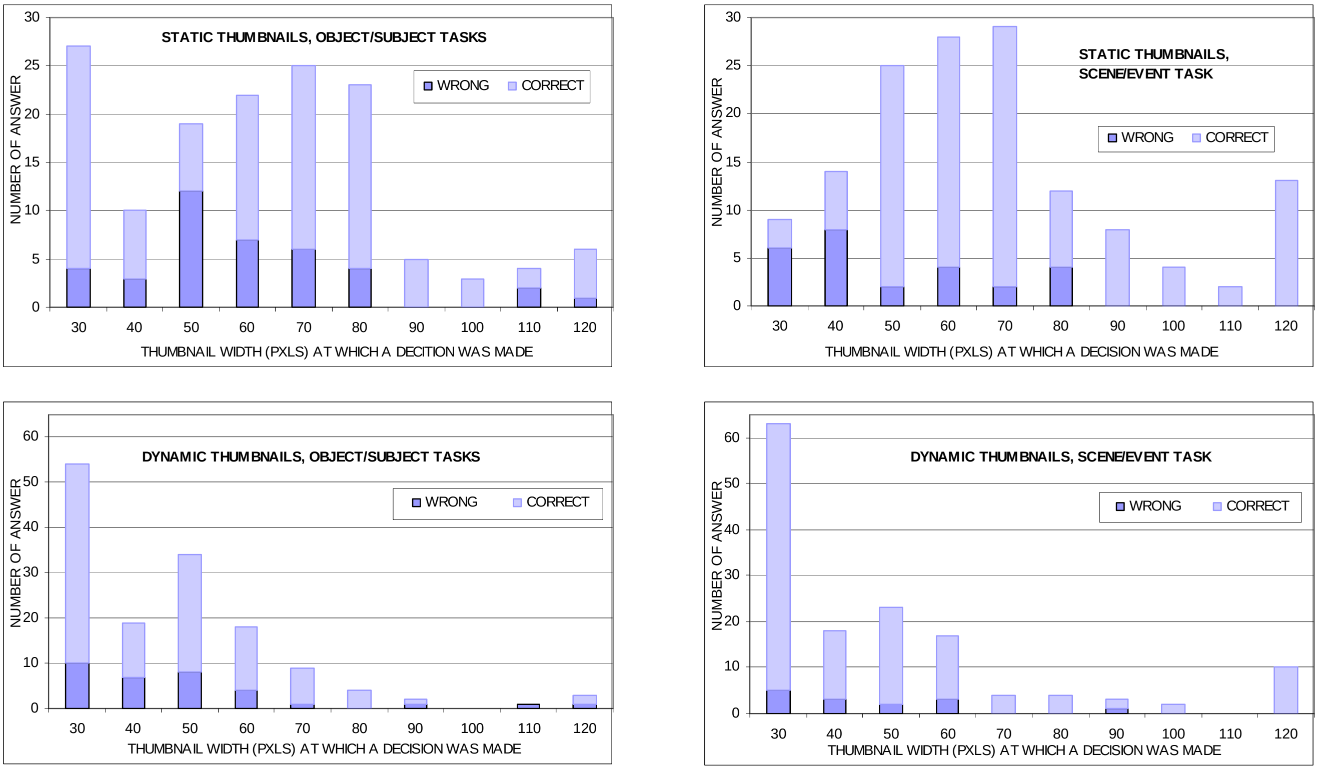


Figure 6. Results for experiment A split into object/subject- (left) versus scene/event-related tasks (right).

rather low (39 vs. 33), the high peak at the smallest thumbnail size for object/subject tasks does not appear for scene/event tasks (cf. top diagrams in Fig. 6), thus confirming our intuitive assumption that dynamic thumbnails perform better for these type of tasks. However, for larger static thumbnail sizes human performance is good even for scene/event tasks, as already indicated above.

Comparing the results of **experiment B** with respect to different task types did not reveal any notable difference compared to the observations based on the general data illustrated in Figure 5.

3. CONCLUSION

In this paper, we quantified the usage of static and dynamic thumbnails for interactive video retrieval on a handheld device. Contrary to widespread believe that screens of handheld devices are unsuited for visualizing multiple (small) thumbnails simultaneously, our results suggest that users are quite able to handle and assess multiple thumbnails, especially when they are showing moving images. This result suggests promising avenues for future research with respect to the design and interaction with advanced video retrieval interfaces on mobile devices. Although the limited screen estate of handheld devices allows for less advanced video retrieval interfaces than those common for the desktop, they can be still much more complex that one would assume, especially when they rely on moving images. Therefore, when designing mobile video retrieval interfaces we recommend keep moving!

Table 1. Average thumbnail sizes (width in pixels)

	OBJECT/SUBJECT TASK			SCENE/EVENT TASKS		
	ALL	CORRECT	WRONG	ALL	CORRECT	WRONG
STATIC	62	63	60	67	71	50
DYNAMIC	47	46	49	49	49	46

4. REFERENCES

[1] C. Gurrin, L. Brenna, D. Zagorodnov, H. Lee, A.F. Smeaton, and D. Johansen. Supporting Mobile Access to Digital Video Archives Without User Queries. In *Proc. MobileHCI*, 2006.

[2] B. Huurnink, C.G.M. Snoek, M. de Rijke, and A.W.M. Smeulders. Today's and Tomorrow's Retrieval Practice in the Audiovisual Archive. In *Proc. ACM CIVR*, 2010.

[3] H. Lee, A.F. Smeaton, N. Murphy, N. O'Connor, and S. Marlow. Fischlar on a PDA: Handheld User Interface Design to a Video Indexing, Browsing, and Playback System. In *Proc. UAHCI*, 2001.

[4] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation Campaigns and TRECVid. In Proc. *ACM Multimedia Information Retrieval*, pp. 321-330, 2006.

[5] A. Torralba, R. Fergus, and W.T. Freeman. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(11):1958-1970, 2008