COMBINING MULTIPLE REPRESENTATIONS ON THE TRECVID SEARCH TASK

Arjen P. de Vries, Thijs Westerveld

CWI
Amsterdam
The Netherlands

*Tzveta Ianeva Ianeva**

University of Valencia
Valencia
Spain

ABSTRACT

This paper[1] presents a (preliminary) analysis of the evaluation results obtained on the TRECVID 2003 search task. We study in particular the effects of combining multiple representations on retrieval: multiple representations of video content (speech and visual) and of the user information need (multiple visual examples). We conclude from our multi-modal retrieval experiments the following *working hypothesis*: even though the ASR run is *usually* better than the visual run, matching against both modalities ensures robustness against choosing the wrong content representation. For the same reason, using multiple visual examples to represent the user information need is preferable over using a single designated example only.

1. INTRODUCTION

Often it is stated that a successful video retrieval system should take advantage of information from all available sources and modalities. Merging knowledge from for example speech, vision, and audio would yield better results than using only one of them.

The TRECVID 2001 and 2002 evaluations have however demonstrated that it is far from trivial to take advantage of multiple representations for video search tasks. Results on automatic speech recognition (ASR) transcripts have been better than any other run, for almost all queries. Similarly, a single visual example query usually outperforms runs based on multiple examples. This year, on the TRECVID 2003 search task, we found that the improvements to the visual modelling in our system have changed this situation: a combination of ASR and visual performs better than either alone. This paper presents a (preliminary) analysis of the evaluation results.

Our retrieval system is based on generative probabilistic models, described briefly in Section 2 and more extensively in [1]. We describe concisely a dynamic variant of the model that allows for describing spatio-temporal information as opposed to the spatial information captured in the basic static models. The dynamic models represent shots instead of still keyframes. Speech transcripts (ASR) are modeled with language models. Section 4 presents experiments combining results from two modalities. Section 5 discusses the usage of multiple (visual) examples.

*The third author performed the work while at CWI, supported by grants GV FPIB01_362 and CTESRR/2003/43.

[1]See http://www.cwi.nl/projects/trecvid/trecvid2003.html for a color version of this document.

2. RETRIEVAL MODEL

The retrieval model for ranking video shots is a generative probabilistic model. Using generative models of textual information is known as the language modelling approach to information retrieval (see e.g. [2]). We have extended it to multimedia retrieval by integrating a (related) probabilistic approach to image retrieval developed in [3]. Details are provided in [1].

The visual model ranks images by their probability of generating the samples (pixel blocks) in one or more query example(s). In the *static model*, keyframe images are modelled as mixtures of Gaussians with a fixed number of components ($C = 8$). The image samples are 8 by 8 pixel blocks, described by their DCT coefficients and their position in the image plane; the models are trained using standard EM [4], assuming a diagonal covariance matrix.

The *dynamic model* is a Gaussian Mixture Model in DCT-space-time domain. It extends the static model with the time dimension (like [5]). Instead of a single image (keyframe), a one-second video sequence around the keyframe is modelled. The sampling process is similar to the one just described. We take 29 frames around the keyframe and cut them in distinct blocks of 8 by 8 pixels. Each block is then described by its DCT coefficients, its x and y position in the image plane and its position in time (normalised between 0 and 1). Given this setup, the static model can be seen as a special case of the dynamic model where the temporal feature takes a fixed value of 0.5. The intuitive explanation in this case is: what happens before and after the central time moment matching the keyframe is unknown.

The training process remains the same: feature vectors are fed to the EM algorithm to find the mixture parameters. Because we use diagonal covariance matrices, components are aligned to the axes. The resulting models capture the (dis-)appearance of objects (but cannot describe temporal events like moving up or down). Figure 1 shows the frames in an example video sequence and a visualisation of the corresponding model. The *tree* in the top left corner is only visible at the beginning of the sequence. The corresponding component in the model also disappears at about $t = 0.5$.

Note that the number of samples for training the dynamic model is much larger than for the static model (29 times as large). While the dynamic models represent some of the temporal events in a shot, other advantages over the static models include more training data describing the visual content, and reduced dependency on choosing an appropriate keyframe.

The *language model* for ASR transcripts represents the video as a sequence of scenes, each consisting of a sequence of shots. The generative model mixes models for shots and scenes with the background collection model. The main idea behind this approach is that a good shot contains the query terms and is part of a scene

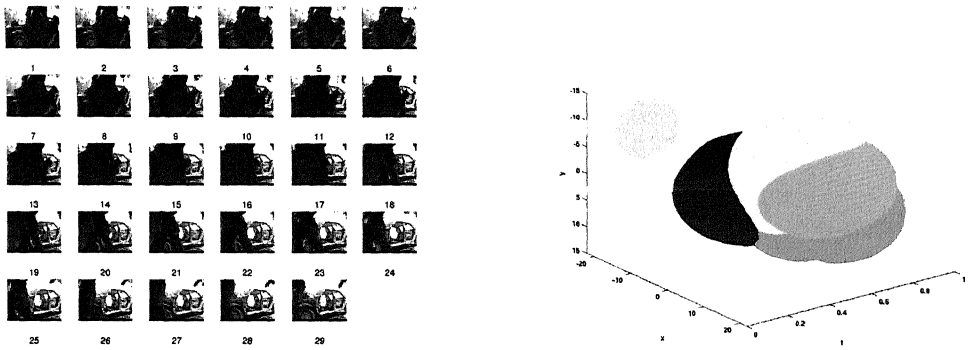

Fig. 1. A shot represented by 29 frames around the keyframe (left) and a 3D visualisation of dynamic GMM computed from it (right). The visualisation plots the Gaussian components with standard deviation below 2 in 3-D $x - y - t$ space, in colour and texture of the mean DCT-coefficients. Prior probabilities and variance in colour and texture are not visualised.

Fig. 2. Visualisation of the relevant items in the ASR run, the RR combined run, and four individual example runs (on the dynamic model alone and the combination with ASR), for the Basketball topic (101).

having more occurrences of the query terms. Also, by including scenes in the ranking function, we hope to retrieve the shot of interest, even if the video's speech describes it just before it begins or just after it is finished. Because scene boundaries are not known, we assume (pragmatically) that each sequence of 5 consecutive shots forms a scene. The samples are the word tokens from the transcript provided by LIMSI (using [6]). Following [2], the foreground probabilities for shots and scenes are estimated using term frequency, and the background probabilities using document frequency. The mixing parameters have been set to the values giving best results on the TRECVID 2002 test collection.[2]

3. EXPERIMENTAL SETUP

The search collection is indexed using the procedures described above. For each shot, we build a static model, a dynamic model, and a language model. Building queries from the topic descriptions is mostly automatic. The only manual action in constructing visual queries has been the selection of one or more image or video examples to be used for ranking. A textual query was constructed manually for each topic, taking only the content words from the topic description.[3] From there on, the whole retrieval process was automatic (except for some of the experiments described in section 5). All image examples are rescaled to at most 272x352 pixels and then JPEG compressed at a quality level of 20%, to match size and quality of the video collection. When a static image is used to query the collection of dynamic shot models, we extend its feature vector with a time value of 0.5; thus assuming nothing is known before or after the moment where it matches the shot's keyframe.

[2] $\lambda_{\text{Shot}} = 0.090$, $\lambda_{\text{Scene}} = 0.210$, and $\lambda_{\text{Coll}} = 0.700$.

[3] Designated examples and textual queries used are available from `http://www.cwi.nl/projects/trecvid/trecvid2003.html`.

4. COMBINING MODALITIES

Assuming (unrealistically) independence between textual and visual information, the joint probability of observing query text and visual example is the multiplication of the individual probabilities (or sum the log probabilities).

Table 1 summarises the results, while Figure 4 shows for each topic the results from each modality. An important result is that it is not needed to choose only one modality; for most queries, the combination is close to the best of the two. A combination of the dynamic models of shots with language models of the ASR transcripts outperforms the individual runs on their own. A combination with the static run performs however worse than ASR only – surprising since the MAP score for the static run is the same as the dynamic MAP. Figure 4 shows that this can be explained as the dynamic run has a higher initial precision. A possible explanation is that the dynamic model does capture some temporal behaviour in the shot. It seems however more plausible that this improved initial precision results from better trained models of visual content, together with reduced dependency on choosing an appropriate keyframe. The sample likelihood computed from the dynamic models thus captures the visual similarity between shots and query example more robustly than the likelihood computed from the static models.

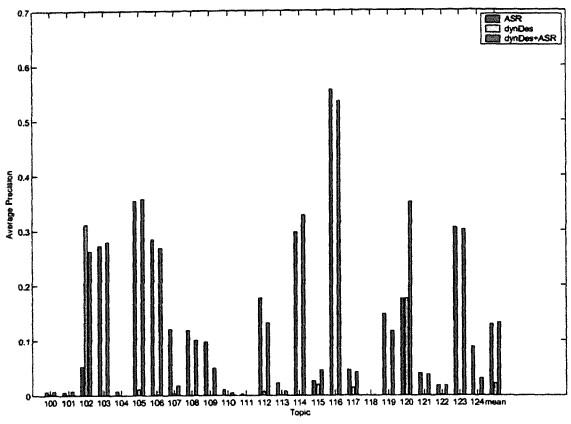

Fig. 3. Visual vs. ASR results (using the designated examples).

Table 1. Mean Average Precision (MAP) for ASR only, visual only and combined runs (static and dynamic models).

		+ASR
no visual		.130
static	.022	.105
dynamic	.022	.132

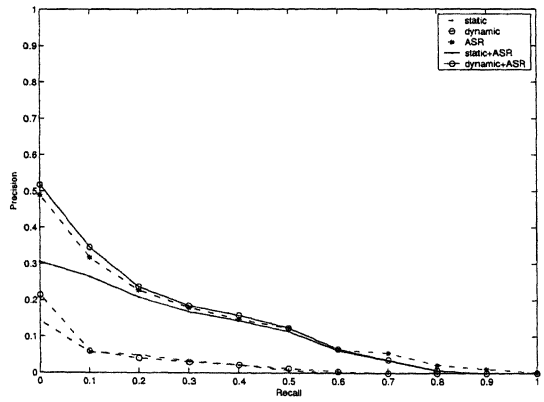

Fig. 4. Recall-Precision graph for static, dynamic and ASR runs as well as multimodal combinations.

Table 2. Mean Average Precision (MAP) for runs with single query, an a priori manual selection, and all query examples; the best single query is an a posteriori result ('cheating')).

	dynamic	+ASR
single	.022	.132
all	.031	.149
selection	.039	.151
best single	.050	.155

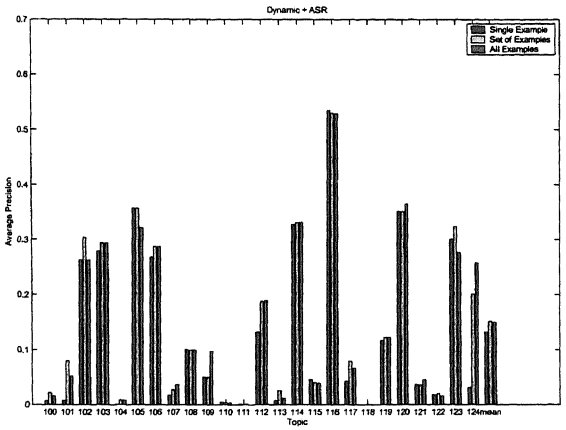

Fig. 5. Results on a single example, a manual selection of examples, and all examples (ASR+Visual).

5. COMBINING TOPIC EXAMPLES

Choosing just one 'best' topic example as a query increases the risk of missing a lot of relevant material (suppose, for example, we selected a close up shot of a point being scored in basketball, and most of the relevant shots in the collection happen to be overview shots of the playing field). This risk may be reduced by representing the user's information need by multiple examples. Table 2 and Figure 5 present the results of an experiment running separate queries for each example and merging run results afterwards. The first bar in the Figure shows the results when the user information need is represented by just the designated example (using both modalities). The second bar is produced as follows. We manually selected a set of 'good' examples (a priori), ran separate queries for each example, and then merged the results using a simple round-robin approach. Duplicates are filtered out afterwards. The result of this process for *all* examples (instead of a selection) is displayed in the third bar. The Figure clearly shows that using a selection of query examples is usually better than restricting ourselves to a single example. Simply using all examples is usually only slightly worse. The 'best single example' row of the Table shows however that choosing a single example could have given better results than our manual selection, had we known how to select the right query images or shots.

6. ANALYSIS OF SPECIFIC TOPICS

Let us investigate some of the results in more detail. Figure 4 shows a small number of exceptions to the general rule that the combination of both modalities is close to the performance of the best of each.

The results of topic 120 ('Dow Jones rising') indicate that the combination of speech and visual can be significantly better than a visual or speech run individually. However, this result is best explained by the specific collection and query examples, that provided the exact same stock-market line graphs as shown on CNN economic news. The ad-hoc topic becomes essentially a known-

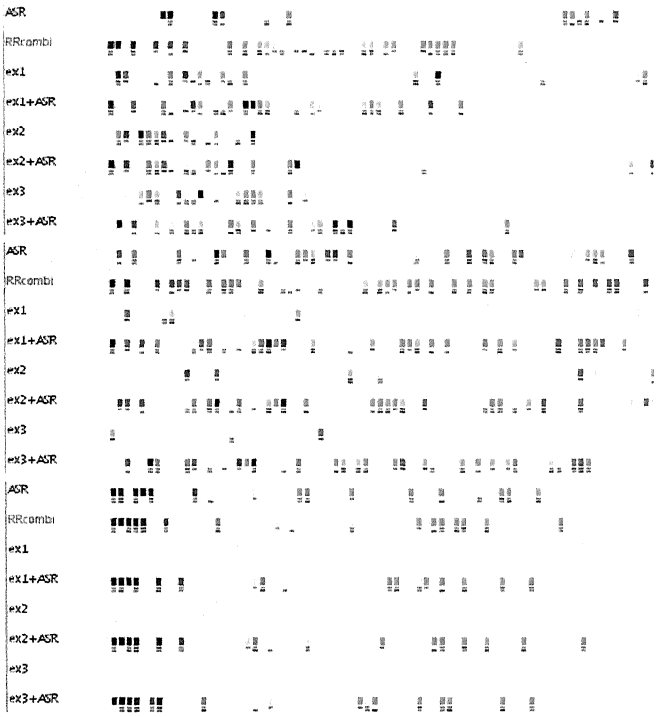

Fig. 6. Visualisations of the relevant items in the ASR run, the RR combined run, and three individual example runs (on the dynamic model alone and the combination with ASR), for topics Yasser Arafat (103), Dow Jones (120), and Flames (112).

item search in the visual modality, simplifying the search task significantly. A quick glance over the top visualisation in Figure 6 (created with NIST's `BeadPlot`[4] tool for visual comparison of TREC result sets) gives away how the relevant results from ASR are pushed up the result list by the visual example runs (especially for the first example, 'ex1').

Now look at the other beadplots in Figure 6. The middle one visualises the runs for three examples of topic 112 ('Flames'): each example contributes for each modality some unique relevant shots. Indeed, Figure 5 confirms that combining all examples of topic 112 is better than using the designated example alone. The bottom beadplot (for topic 103, 'Yassar Arafat') shows an example where the visual results do not contribute any relevant results, but do not harm the ASR results either: the combined run still contains most relevant examples from the speech run in the top results.

Unfortunately, Figure 2 (a beadplot for the 'basketball' topic) demonstrates that this is not always the case. For this topic, the visual context coincides with the user information need: query example three retrieves many relevant items but the combination with ASR moves these results far down the ranked list (degrading the MAP score significantly). The query is similar in type to the baseball topic (102), for which visual results are much better than the ASR as well; here, the combination of speech and visual did however not degrade the results significantly.

[4] See `http://www.itl.nist.gov/iaui/894.02/projects/beadplot/`.

7. CONCLUSIONS

This paper analysed the effects of combining multiple representations of both video content (multiple modalities) and user information need (multiple visual examples) on the TRECVID 2003 search task. Averaged over all twenty-five queries, it is better to rank on multiple modalities. We derive the following *working hypothesis*: matching against both modalities gives robustness.

The average precision of the individual visual results are comparable, and in most cases neither dynamic nor static models are good representations of the user information needs. The differences in results when combined with ASR results indicate however that the dynamic shot models capture better the similarity to visual example items. Combination with the keyframe representation of a shot (i.e., the static model) is too fragile.

We conclude from the experiments using multiple query examples that the visual aspects of the user information need are best represented by multiple visual examples. Manual selection of 'good' visual examples for a given topic gave the best results among the visual runs. Ranking (multiple) good examples on both visual and speech modalities, followed by combining these per-example runs in round-robin fashion, gives near-best results for almost all topics.

Further result analysis has to be performed to better understand the retrieval results. A relatively straightforward experiment is to check whether the conclusion about using a combination of speech and visual still holds after we improve our ASR runs to the level of average precision reported by CMU on the TRECVID 2003 search task. More research is needed to find out which assumption explains best the observed advantages of the dynamic model over the static one: more training data, less dependency on the keyframe, or the spatio-temporal aspects of the model.

8. REFERENCES

[1] Thijs Westerveld, Arjen P. de Vries, Alex van Ballegooij, Fransicska M. G. de Jong, and Djoerd Hiemstra, "A probabilistic multimedia retrieval model and its evaluation," *EURASIP Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 186–198, 2003.

[2] Djoerd Hiemstra, *Using language models for information retrieval*, Ph.D. thesis, Centre for Telematics and Information Technology, University of Twente, 2001.

[3] Nuno Vasconcelos, *Bayesian Models for Visual Information Retrieval*, Ph.D. thesis, Massachusetts Institut of Technology. 2000.

[4] A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, series B*, vol. 39, no. 1, pp. 1–38, 1977.

[5] Hayit Greenspan, Jacob Goldberger, and Arnaldo Mayer, "A probabilistic framework for spatio-temporal video representation and indexing," in *European Conference on Computer Vision (ECCV)*, 2002.

[6] J.L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, no. 1–2, pp. 89–108, 2002.