# Measuring the Influence of Concept Detection on Video Retrieval

Pablo Toharia[1], Oscar D. Robles[1], Alan F. Smeaton[2], and Ángel Rodríguez[3]

[1] Dpto. de Arquitectura y Tecnología de Computadores, Ciencias de la Computación
e Inteligencia Artificial,
U. Rey Juan Carlos, C/ Tulipán, s/n. 28933 Móstoles. Madrid. Spain.
{pablo.toharia,oscardavid.robles}@urjc.es,
[2] CLARITY: Center for Sensor Web Technologies, Dublin City University,
Glasnevin, Dublin 9, Ireland
alan.smeaton@dcu.ie
[3] Dpto. de Tecnología Fotónica, U. Politécnica de Madrid,
Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain
arodri@fi.upm.es

**Abstract.** There is an increasing emphasis on including semantic concept detection as part of video retrieval. This represents a modality for retrieval quite different from metadata-based and keyframe similarity-based approaches. One of the premises on which the success of this is based, is that good quality detection is available in order to guarantee retrieval quality. But how good does the feature detection actually need to be? Is it possible to achieve good retrieval quality, even with poor quality concept detection and if so then what is the "tipping point" below which detection accuracy proves not to be beneficial? In this paper we explore this question using a collection of rushes video where we artificially vary the quality of detection of semantic features and we study the impact on the resulting retrieval. Our results show that the impact of improving or degrading performance of concept detectors is not directly reflected as retrieval performance and this raises interesting questions about how accurate concept detection really needs to be.

## 1 Introduction and Background

The automatic detection of semantic concepts from video is opening up a completely new modality for supporting content-based operations like search, summarisation, and directed browsing. This approach to managing content compliments using video metadata and using keyframe similarity and is being enabled by improvements in the accuracy, and the number of, such detectors or classifiers by many research groups. This can be seen in the recent development in activities such as TRECVid [1] where it is now realised that retrieval systems based on low-level features like colour and texture do not succeed in describing high-level concepts as a human would do.

Various authors are now making efforts on optimizing automatic detection of semantic concepts for use in applications such as retrieval. However, it is not clear what is the real impact of improving the accuracy of the detection process, i.e. whether a significant improvement in the performance of detection will yield better quality retrieval. There have been some previous studies of the efficiency of using concepts in retrieval [2–4]. Recently, Snoek et al. [5] analyzed whether increasing the number of concept detectors as well as their combination would improve the performance of retrieval and found that it does.

Wang and Hua study how to improve the performance of combining video concept detectors when dealing with a large number of them by following a Bottom-Up Incremental Fusion (BUIF) approach [6], but they do not deal with the issue of assessing detectors' real influence in retrieval. Thus it appears there is no work studying the relationship between the quality of detectors and retrieval performance. The work here explores the relationship between concept detection performance and content-based retrieval and to examine whether improving detection will yield an improvement at the retrieval stage, or whether this is worth the effort.

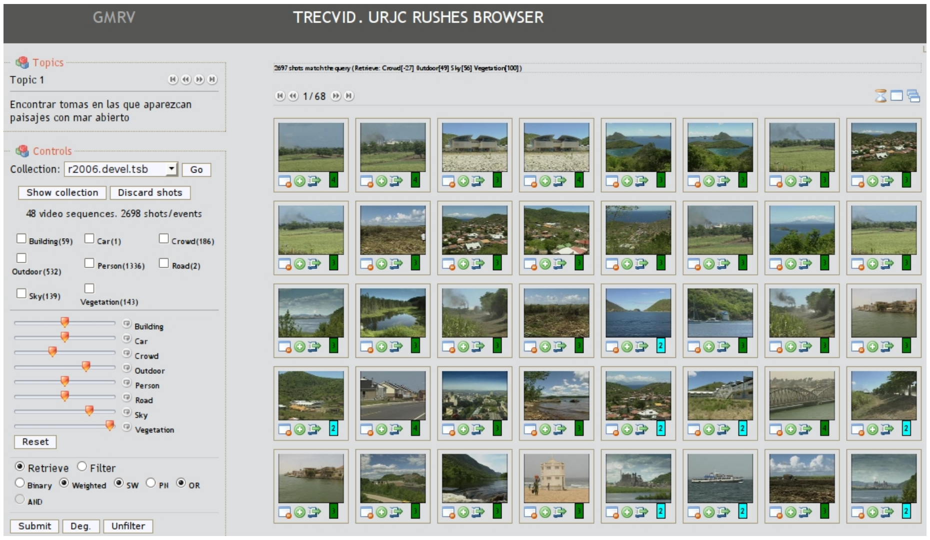
## 2 Materials and Methods

We now detail how we set up an experimental environment for video retrieval using semantic concepts. Controlled noise in concept detection is introduced so as to improve or worsen it, allowing performance of retrieval to be measured. Section 3 presents experiments together with the analysis and conclusions reached.

### 2.1 Concept Detection

The first step is to set up a system to extract concepts from shots. In our work we used the TRECVid [7] 2006 rushes collection of 27 hours which gave rise to approximately 2,900 shots. The concepts selected to work with are defined from within LSCOM-Lite, a reduced version of the 449 Large Scale Concept Ontology for Multimedia [8] annotated concepts that form LSCOM.

The concept detection process is broken into several steps. First, a preprocessing stage extracts keyframes that represent the video content. These are then filtered in order to discard shots such as calibration charts, black frames and so on. We then extract low-level features for these keyframes which are then used as the input to the 39 classifiers. More details of the keyframe extraction and filtering stages can be found in [9]. Finally, Support Vector Machines (SVM) provided by Dublin City University from our high level feature detection submission in TRECVid 2006 are used, using low-level primitive features like colour and texture, extracted by the AceToolbox [10]. The concept classifiers each provide a certainty value $C_i \in [-1, 1]$ that each of the shots' keyframes in the original video contains each of the concepts and we use these as baseline examples of the accuracy of a real implementation of concept detection.



Topic 1

Encontrar tomas en las que aparezcan paisajes con mar abierto

Collection: r2006.devel.tsb

48 video sequences. 2698 shots/events

□ Building(59) □ Car(1) □ Crowd(186)
□ Outdoor(532) □ Person(1336) □ Road(2)
□ Sky(139) □ Vegetation(143)

● Retrieve ○ Filter
○ Binary ● Weighted ● SW ○ PN ● OR
○ AND

2697 shots match the query (Retrieve: Crowd[-27] Outdoor[49] Sky[56] Vegetation[100])

1/68

Fig. 1: Retrieval example using weighted concepts.

### 2.2 Interactive concept-based retrieval engine

An interactive video retrieval system is used in order to test the relationship between the quality of detected concepts and retrieval performance. This allows a user to select which of the available concepts should be used in retrieval, as well as fixing $W_i$ weights for each of the concepts. These are positive if the concept is relevant for the query, and if its absence is relevant to the query it will be negative, else it will be 0. The retrieval engine will assign a value $score_i$ for each shot so a sorted list of results can be presented to the user. Assuming there are $N$ concepts the following is how we obtain a score for each shot:

$$shot_i = \{C_{i1}, C_{i2}, \ldots, C_{iN}\}\ , \quad C_{ij} \in [-1, 1] \tag{1}$$

$$score_i = \frac{\sum_{i=1}^{N} W_i \cdot C_i}{N}\ , \quad W_{ij} \in [-1, 1] \tag{2}$$

As was previously stated, other approaches to combining concept features in retrieval are possible, such as proposed by Wang and Hua [6] or by Snoek and Worring [5], but in our present work we were not interested in addressing the detector fusion method. Figure 1 shows a retrieval result based on 8 concepts selected by the user. On the left side, 8 sliding bars allow a user to adjust weights for each concept and a visualization of the top-ranked shots is also shown.

### 2.3 Degradation and improvement of concept detection

Performing an artificial degradation or improvement of concept detection quality can be achieved by introducing noise into the concept detector output, so the certainty value is increased or decreased as needed. However, rather than depend on the accuracy of automatic detection only, the existence of a ground truth allows us to faithfully simulate improvement and degradation of concept

**Table 1:** Concepts used in experiments.

| Concept | Description |
|---|---|
| Building | Shots of an exterior of a building |
| Car | Shots of a car |
| Crowd | Shots depicting a crowd |
| Outdoor | Shots of Outdoor locations |
| Person | Shots depicting a person. The face may be partially visible |
| Road | Shots depicting a road |
| Sky | Shots depicting sky |
| Vegetation | Shots depicting natural or artificial greenery, vegetation woods, etc. |

detection. To obtain this, a manual process of double annotation of each of the concepts over the whole collection was performed.

To vary detection quality, a percentage $P$ of shots from the collection are randomly selected and their certainty degree is modified for each detector. To improve performance, a value $A$ is added to the certainty value of shots from the ground truth in which the concept is known to be present. If a shot does not contain the concept, the value $A$ will be subtracted from the certainty value. In case of degrading the detectors' performance, the process is reversed.

In measuring the impact of concept detection on retrieval, we use an offline retrieval proces. We use the keyframes of the shots selected to initiate a low-level retrieval process using the low-level image characteristics used as input to concept recognition, to perform keyframe similarity. This generates a content-based ranking of shots for each topic. A concept-based retrieval ranking is also generated using the weights selected by the users and degrading/upgrading the performance of the detectors accordingly. The results of both retrieval rankings are normalized and combined in a 50:50 ratio to give the final retrieval output. While this may seem like diluting the impact of concept retrieval, and concept detection accuracy, it reflects the true way in which video retrieval is carried out in practice. Retrieval performance is evaluated using Mean Average Precision (MAP) over the set of topics and thus by varying the parameters $A$ and $P$, a change in retrieval MAP should be obtained for each concept.

For our experiments we concentrated on a subset of concepts from LSCOM-Lite, shown in Table 1, chosen because they occur throughout the whole video dataset whereas others occur much less frequently. Our experiments are carried out in two parts, an online part working with non-expert users who perform iterative retrieval, and an automated offline part using results from the user retrieval and performing more exhaustive tests varying concept detection quality. This is shown in Figure 2.

### 2.4 Experimental methodology

Our experimental methodology is as follows. In the first stage a user searches for shots given a topic using the interactive system and the concept-based retrieval engine described earlier. Topics have been constructed in such a way that they require iterations of the retrieval system to refine and adjust topic weights until they are optimal, and that they use concepts both in a positive or negative way. Topics are shown in Table 2, along with their number of relevant shots.

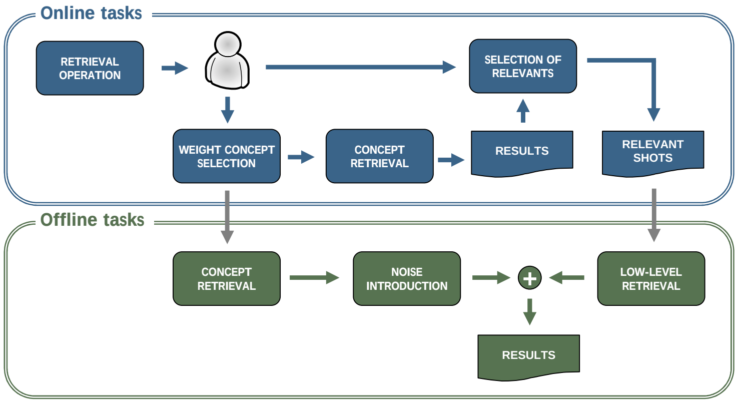


**Fig. 2:** Experimental framework.

**Table 2:** Search topics.

| **Topic** | **Description: "Find shots containing . . . "** | # rel. shots |
|---|---|---|
| 1 | . . . open-sea views | 33 |
| 2 | . . . 2 or more people with plants in an urban area | 243 |
| 3 | . . . desert-like landscapes | 55 |
| 4 | . . . village settlements on the coast | 73 |
| 5 | . . . 2 or more people interacting in a natural environment | 91 |
| 6 | . . . a person talking to an audience inside a building | 39 |
| 7 | . . . people sailing | 42 |

**Table 3:** Use of concepts in Topics.

| | Building | Car | Crowd | Outdoor | Person | Road | Sky | Vegetation |
|---|---|---|---|---|---|---|---|---|
| Topic 1: | 6(-) | 6(-) | 3(-) | 9(+) | 2(-) | 5(-) | 7(+) | 5(-) |
| Topic 2: | 9(+) | 4(+) | 9(+) | 2(+) | 7(+) | 6(+) | 2(-) | 9(+) |
| Topic 3: | 7(-) | 5(-) | 3(-) | 9(+) | 3(+) | 4(-) | 9(+) | 9(-) |
| Topic 4: | 8(+) | 3(+) | 2(+) | 9(+) | 3(+) | 5(-) | 8(+) | 5(+) |
| Topic 5: | 5(-) | 3(-) | 8(+) | 9(+) | 8(+) | 3(-) | 4(+) | 8(+) |
| Topic 6: | 5(+) | 4(-) | 9(+) | 9(-) | 7(+) | 6(-) | 6(-) | 3(-) |
| Topic 7: | 6(-) | 5(-) | 5(+) | 9(+) | 8(+) | 7(-) | 9(+) | 5(+) |

Topics will use available concepts in positive or negative ways, depending on the subject matter. Topic 6 can be associated with negative weighting of the concept "Outdoor", since the aim is that action takes place inside a building. Table 3 shows the ways that the set of 9 users use topics in positive or negative ways. For example for Topic 2, 4 users used the concept "Car" in a positive way and 2 used "Sky" in a negative way. Some aspects of some topics may not be addressable in query formulation with the available concepts and while this may seem a limiting factor, it is also representative of a real world search where there will never be enough appropriate concepts for the variety of user search topics.

For our experiments, 9 users without any professional experience of searching were recruited. Each user was given an introduction to the system, and the 7 topics were presented in a rotating order to avoid bias. Each user adjusted concept weights for each of the topics and a retrieval operation was performed with

**Table 4:** MAP variation average for retrieval introducing controlled noise into detector performance.

(a) Degradation

| P/A | -0.1 | -0.3 | -0.5 |
|---|---|---|---|
| 10% | -0.10% | -0.74% | -1.23% |
| 30% | -0.67% | -2.61% | -4.25% |
| 50% | -0.92% | -3.42% | **-5.69%** |

(b) Improvement

| P/A | +0.1 | +0.3 | +0.5 |
|---|---|---|---|
| 10% | 0.10% | 0.19% | 0.53% |
| 30% | 0.45% | 1.20% | 2.16% |
| 50% | 0.81% | 2.12% | 3.98% |

the user marking relevant shots or adjusting concept weights and performing a new search. Once the sets of relevant shots had been identified, we can calculate retrieval rankings based on combined weighted concept-based and content-based techniques and calculate MAP retrieval performance by measuring against an exhaustive manual assessment of shot relevance, our ground truth for retrieval. We can examine the effect of detector quality on retrieval performance by introducing noise into the output of the concept detectors as described in section 2.3. Each variation on the parameters that results in degraded or improved detectors gives a new list of ranked shots which can be evaluated against the ground truth, and MAP calculated. Combining the different options available, we have a total of 9 users, each running 7 queries with improvements and degradations on 8 concepts, to be evaluated.

## 3 Results and Discussion

### 3.1 Performance of retrieval

Table 4 shows the average MAP percentage variations when we degrade or improve the quality of the underlying concept detection above or below the level of concept detection performance obtained from the automatic DCU concept detection. Thus we use the real performance figures as a baseline and vary detection quality above and below this. The MAP performance using unmodified concept detection performance is 0.0254.

What these results tell us, for example, is that when we degrade concept detection performance for all concepts by reducing the certainty value for detection by 0.5 (on a scale of -1 to 1) for 50% of the shots, we get a net drop in MAP performance for retrieval of only 5.69% (bolded entry in Table 4).

Table 5 collects the average Coefficient of Variation values considering the results achieved by all users and among all topics. Coefficient of Variation values are more stable across users rather than across topics, but the worst cases appear for the lower values of $P$ and $A$ variables because the average variations are very low (Table 4). This can be due to the user interaction with the retrieval engine and to the random controlled noise introduced.

### 3.2 Detector performance versus the retrieval task

Figure 3 shows MAP variations when fixing one of the parameters, either $A$ or $P$, for both detection performance and for concept retrieval performance. The x-axis depicts $A$ or $P$ values for improvement (represented as positive values of the scale) or degradation (negative values). The y-axis shows the variation

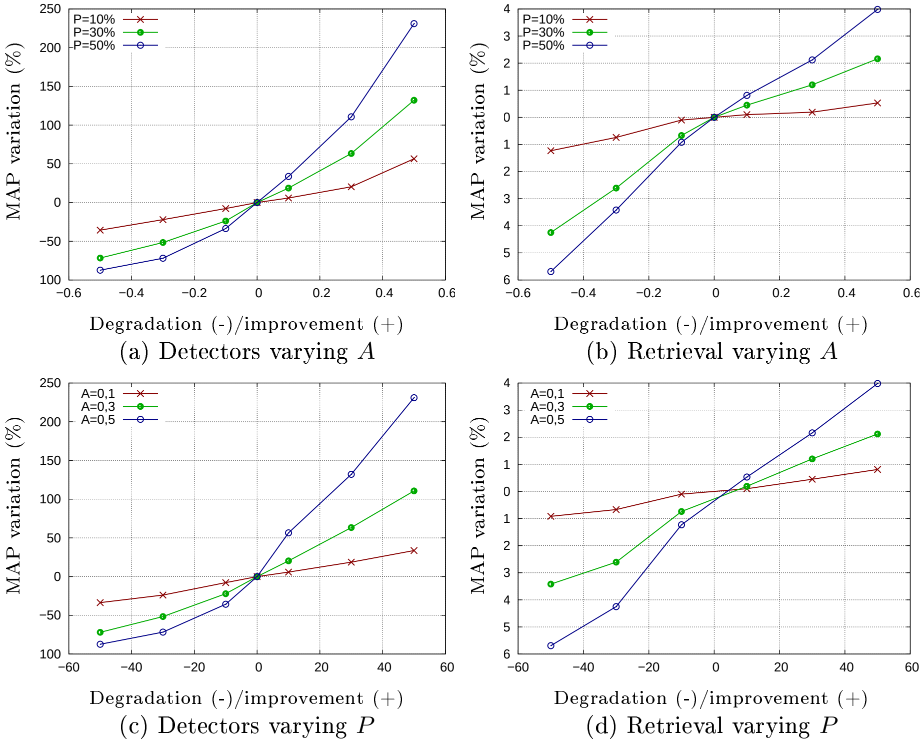
**Table 5:** Avg. Coefficients of Variation considering responses by users, all topics.

(a) Per user.

| Degradation | | | | Improvement | | | |
|---|---|---|---|---|---|---|---|
| P/A | -0.1 | -0.3 | -0.5 | P/A | +0.1 | +0.3 | +0.5 |
| 10% | 5.277 | 0.638 | 0.723 | 10% | 2.454 | 8.337 | 5.078 |
| 30% | 1.156 | 0.327 | 0.350 | 30% | 1.960 | 1.764 | 1.285 |
| 50% | 1.299 | 0.453 | 0.414 | 50% | 1.081 | 1.105 | 0.929 |

(b) Per topic.

| Degradation | | | | Improvement | | | |
|---|---|---|---|---|---|---|---|
| P/A | -0.1 | -0.3 | -0.5 | P/A | +0.1 | +0.3 | +0.5 |
| 10% | 8.361 | 1.794 | 1.835 | 10% | 6.052 | 12.117 | 7.205 |
| 30% | 1.881 | 1.182 | 1.231 | 30% | 2.511 | 2.523 | 2.043 |
| 50% | 1.620 | 1.195 | 1.214 | 50% | 1.372 | 1.276 | 0.986 |



(a) Detectors varying $A$

(b) Retrieval varying $A$

(c) Detectors varying $P$

(d) Retrieval varying $P$

**Fig. 3:** MAP variations for detection and retrieval, varying $A$ and $P$.

of the MAP in percentages. The curves show similar trends for both $A$ and $P$ transformations for all the tests shown. However, Figures 3(a) and 3(b) ($A$ as parameter) show different range values for positive (improvement) and negative (degradation) intervals, being the variation most noticeable in the improvement transformation. On the other hand, both the tendency and the interval of Figures 3(c) and 3(d) ($P$ as parameter) are very similar. Overall, however, we can say that the impact of detection accuracy is far less pronounced than we would expect, indicating that even poor detection accuracy provides useful retrieval.

## 4 Conclusions

We have implemented a methodology to analyze the impact of concept detection accuracy on video retrieval on a collection of rushes video. We found that even poor quality detection can yield good retrieval and that as the quality of detection improves the quality of retrieval does not rise accordingly. While this may appear as just an interesting exercise and the results do depend on the set of concepts used, it does represent the state of the art in using concepts in retrieval, as shown in TRECVid, where it is shown that exploiting the dependencies among concepts is non-existent. For future work we plan to further investigate how detection performance is impacted when semantic dependencies among concepts (e.g. "`Outdoor/Building`" and "`Person/Crowd`") and this will integrate concept ontologies into our work. Other work will be to extend the number of concepts to see if similar results are obtained for concepts which do not occur as frequently in the video as the ones used here.

**Acknowledgments** This work has been partially funded by the the Spanish Ministry of Education and Science (grant TIN2007-67188). BBC video was provided by TRECVid [1]. AS was funded by Science Foundation Ireland under grant 07/CE/I1147.

## References

1. Smeaton, A.F., Over, P., Kraaij, W.: High-Level Feature Detection from Video in TRECVid: a 5-Year Retrospective of Achievements. In Divakaran, A., ed.: Multimedia Content Analysis, Theory and Applic. Springer Verlag, Berlin (2009) 151–174
2. Hauptmann, A.G., Yan, R., Lin, W.H., Christel, M.G., Wactlar, H.D.: Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. IEEE Transactions on Multimedia **9**(5) (2007) 958–966
3. Christel, M.G., Hauptmann, A.G.: The use and utility of high-level semantic features in video retrieval. In Leow, W.K., Lew, M.S., Chua, T.S., Ma, W.Y., Chaisorn, L., Bakker, E.M., eds.: CIVR. Volume 3568 of LNCS, Springer (2005) 134–144
4. Wei, X.Y., Ngo, C.W.: Fusing semantics, observability, reliability and diversity of concept detectors for video search. In: Proc. MM '08, NY, USA, ACM (2008) 81–90
5. Snoek, C.G.M., Worring, M.: Are concept detector lexicons effective for video search? In: ICME, IEEE (2007) 1966–1969
6. Wang, M., Hua, X.S.: Study on the combination of video concept detectors. In: Proc. MM '08 NY, USA, ACM (2008) 647–650
7. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proc. MIR '06, NY, USA, ACM Press (2006) 321–330
8. Naphade, M.R., Kennedy, L., Kender, J.R., Chang, S.F., Smith, J.R., Over, P., Hauptmann, A.: A light scale concept ontology for multimedia understanding for TRECVID 2005. Tech. Rep. RC23612, IBM T.J. Watson Research Center (2005)
9. Toharia, P., Robles, O.D., Pastor, L., Rodríguez, A.: Combining activity and temporal coherence with low-level information for summarization of video rushes. In: Proc. TVS '08: NY, USA, ACM (Oct. 2008) 70–74
10. O'Connor, N., Cooke, E., le Borgne, H., Blighe, M., Adamek., T.: The acetoolbox: Low-level audiovisual feature extraction for retrieval and classification. In Proc. EWIMT'05 (2005)