ORAND at TRECVID 2015: Instance Search and Video Hyperlinking Tasks

Juan Manuel Barrios
ORAND S.A.
Santiago, Chile
juan.barrios@orand.cl

Jose M. Saavedra
ORAND S.A.
Santiago, Chile
jose.saavedra@orand.cl

Felipe Ramirez
ORAND S.A.
Santiago, Chile
felipe.ramirez@orand.cl

David Contreras
ORAND S.A.
Santiago, Chile
david.contreras@orand.cl

ABSTRACT

ORAND S.A. is a Chilean company focused on developing applied research in Computer Science. This report describes the participation of the ORAND team at Instance Search (INS) and Video Hyperlinking (LNK) tasks in TRECVID 2015.

The INS participation consisted in four submissions to automatic detection. All the submissions used the four samples for each topic without using video queries (type A). We tested different score propagation algorithms where the based on low-level features achieved best performance.

The LNK participation consisted in four submissions. The submissions considering low-level features (color histogram, edges histogram and acoustic energies) outperformed the semantic descriptors (concepts and captions).

1. INTRODUCTION

ORAND is a Chilean software company focused on developing applied research in Computer Science. This paper describes our participation at Instance Search (INS) and Video Hyperlinking (LNK) tasks at TRECVID 2015 [10]. TRECVID is an evaluation sponsored by the National Institute of Standards and Technology (NIST) with the goal of encouraging research in video information retrieval [11].

2. INSTANCE SEARCH

Instance Search task (INS) consists in retrieving the shots that contain a given entity (object or person) from a video collection. The target entity, called a *topic*, is defined by visual examples and a brief textual description. A visual example is a still image (extracted from a sample video) and a mask, which delimits the region of the image where the topic is visible. INS 2015 evaluated 30 topics (21 unique objects, 5 generic objects, 2 locations and 2 characters) with up to four visual examples per topic [10]. The reference video collection is the same since INS 2013, the BBC EastEnders collection, which consists in 244 videos with a total extension of 435 hours (39 million frames approx.). Additionally, the list of shots for each video was predefined and given to each team (a total number of 471,526 shots). Each participant system had to submit the list of shots that most probably show each topic (with a maximum length of 1000 shots per topic).

2.1 System Description

This participation is the progression of our work in this dataset since TRECVID 2013 [6, 7]. We are currently interested in studying the propagation of scores between similar shots given a baseline of detected shots. Unlike the codebook approach, for computing our baseline we follow the k-NN approach on the full set of descriptors.

As a general overview, the baseline follows these steps: video frames are sampled at a regular-step and local descriptors are computed for the selected frames. The extracted local descriptors are partitioned into subsets, and for each subset a k-NN search is performed. The partial results for all subsets are merged in order to determine the actual k-NN. A voting algorithm ranks each shot according to the number of nearest neighbors they contain in the k-NN lists. Once the detection score is computed for each shot and for each topic, we use different methods for propagating scores between shots according to a pre-computed similarity shot graph.

2.1.1 Feature Extraction

The videos in the collection are TV quality: 576i/25. In particular, interlaced videos show unnatural horizontal lines that may affect the quality of local descriptors. In order to reduce this effect, all the videos were re-encoded and deinterlaced using FFmpeg software [1]. Then, every video was sampled at three frames per second, and for each frame we computed CSIFT implemented by *FeatureSpace* software [3].

2.1.2 Similarity Search

The similarity search consisted in retrieving for each x in $\mathcal{Q}$ the k Nearest Neighbors in $\mathcal{R}$ according to distance:

$$L_1(\vec{x},\vec{y}) = \sum_{i=0}^{d} |x_i - y_i|$$

In order to solve these searches, we partitioned $\mathcal{R}$ into several subsets $\{\mathcal{R}_1, ..., \mathcal{R}_n\}$, i.e.:

$$\mathcal{R} = \bigcup_{i=1}^{n} \mathcal{R}_i \quad , \quad \forall\, i \neq j, \mathcal{R}_i \bigcap \mathcal{R}_j = \emptyset$$

Run	Features	MAP	P_5	P_{10}	P_{20}
L_1_O_M_VO	Subtitles+Concepts+Color+Edge+Audio	0.0845	0.3515	0.2545	0.1712
L_2_F_M_N	Subtitles	0.0377	0.2286	0.1536	0.0857
L_3_F_N_V	Concepts	0.0581	0.3132	0.2075	0.1264
L_4_F_N_VO	Color+Edge+Audio	0.1071	0.4632	0.3132	0.2039

Figure 1: Video Hyperlinking results achieved by our four submissions.

Thereafter, for each x in $\mathcal{Q}$ an approximate k-NN search is performed at every $\mathcal{R}_i$. The final k-NN are determined by merging the n partial results and selecting the top k. The similarity search was implemented using the FLANN library [9].

2.1.3 Voting algorithm

In order to score shots, a voting algorithm traverses the lists of k-NN for each local descriptor at each example image, and sums one vote to the shot that contains the frame that produced the NN. Each votes is weighted according to the distance to the mask and the rank in the k-NN list of the voter. The sum of votes produces the final score for each shot, and the top 1000 are selected for each topic.

2.1.4 Similarity Shot Graph

Given the set S of n shots $S = \{s_1, ..., s_n\}$, the Similarity Shot Graph (SSG) is defined as a weighted graph where the set of nodes corresponds to S and the edge weights $w(s_i, s_j)$ represents the degree of similarity between shots s_i and s_j. The weights are normalized to the range $[0, 1]$, where zero means no similarity and the unitary weight means high similarity [5]. The SSG is created by computing the similarity between every pair of shots in the collection. We implemented an efficient similarity search using the MetricKnn library [2].

Usually, a shot division produces fine-grained segmentation of a scene. In fact, the shot division provided by NIST produces shots with average length 3.3 seconds, and many shots are just a few milliseconds length. Hence, if a static object is visible in a shot, it may be expected that the object will also be visible in other shots from the same scene. In our participation we compared three criteria to construct a SSG:

- low-level global visual similarity, where the degree of dissimilarity is the average distance between three frames per shot (*start*/*middle*/*end*). The distance is the L_1 distance between edge histogram descriptors.
- speech similarity, where the degree of dissimilarity is measured by the cosine distance between vectors summarizing all the words spoken inside shot boundaries (according to BBC captions), following the *tf-idf* model.
- concepts similarity, where the degree of dissimilarity is measured by the cosine distance between vectors summarizing all the concepts detected in shot frames, following the *tf-idf* model. The concepts were detected using the Caffe framework with the pre-trained model AlexNet [8] on sampled frames for each shot. This criterium is analogous to speech similarity but replacing spoken words by detected concepts.

2.2 Submissions and Results

All our submissions were type A (four visual examples, no videos). Each submission was evaluated by NIST, computing the average precision by topic and the mean average precision (MAP). We submitted four automatic runs:

- F_A_1: CSIFT descriptors extracted every 3 frames per second, approximate 100-NN search using 10 kdtrees. MAP=0.223.
- F_A_2: The F_A_1 baseline plus score propagation according to visual similarity between frame shots. MAP=0.229.
- F_A_3: The F_A_1 baseline plus score propagation according to similarity between detected objects. MAP=0.219.
- F_A_4: The F_A_1 baseline plus score propagation according to similarity between words in speech transcription. MAP=0.215.

3. VIDEO HYPERLINKING

Video Hyperlinking (LNK) task consists in retrieving the most relevant video segments for the given anchors. The video dataset correspond to the BBC television broadcasts during almost three months, total 2,686 hours, while the anchors correspond to 100 video segments with 72 seconds length on average.

For our participation we computed different content-based features for each video:

- Color histogram, which correspond to a 512-bins histogram in the RGB space for a 2x2 partition (a 2048-d vector every 0.33 seconds).
- Edge histogram, which correspond to a 10 bins gradient orientation for a 4x4 partition (a 160-d vector every 0.33 seconds).
- Audio frequencies descriptor, which represents the Mel scale in 160 bands, as described in [4] (a 160-d vector every 0.33 seconds).
- Subtitles tf-idf vector, which corresponds to aggregate all the words in the subtitles provided by BBC captions.
- Concepts tf-idf vector, which corresponds to aggregate all the concepts detected in frames according to the detection lists provided by Leuven University [12].

Those descriptors were combined in a late fusion approach to produce the final submission runs. The obtained results for each submissions are described in Figure 1. The best result was obtained by combining visual and acoustic low-level features. Note that when combining the low-level features with semantic features (concepts or subtitles) the performance decreases. This effect was unexpected because concepts or subtitles provides complementary information to low level information. We need more work in order to cor-

rectly fuse those modalities without decreasing the performance.

4. CONCLUSIONS

In this report we described our submissions to INS and LNK tasks at TRECVID 2015. Both were based on resolving k-NN searches in the full set of descriptors without applying quantization. In the Instance Search task, the similarity shot graph may be useful to improve the MAP, but we should note in some topics it may harm the precision. In Video Hyperlinking task, our submissions achieved in general a low performance. We still need more work in order to achieve a satisfactory fusion between low-level features and semantic features.

All the submissions were completed on a single machine Intel Core i7-4770K (3.50GHz, 8 cores), 32 GB RAM, 7 TB disk, Linux.

5. ACKNOWLEDGEMENTS

This research was partially supported by CONICYT Project PAI-78120426.

6. REFERENCES

[1] FFmpeg. `http://www.ffmpeg.org/`.

[2] MetricKnn. `http://www.metricknn.org/`.

[3] Feature Detectors and Descriptors: The State Of The Art and Beyond. Feature Detection Code., 2010. `http://kahlan.eps.surrey.ac.uk/featurespace/web/`.

[4] J. M. Barrios. Content-based video copy detection. In *PhD Thesis*, 2013. https://dx.doi.org/10.13140/2.1.1766.9125.

[5] J. M. Barrios and J. M. Saavedra. Score propagation based on similarity shot graph for improving visual object retrieval. In *Proc. of the Speech, Language and Audio in Multimedia Workshop (SLAM) at ACM Multimedia 2015*. ACM, 2015. http://dx.doi.org/10.1145/2802558.2814644.

[6] J. M. Barrios, J. M. Saavedra, F. Ramirez, and D. Contreras. Orand team: Instance search and multimedia event detection using k-nn searches. In *Proc. of TRECVID*. NIST, USA, 2013.

[7] J. M. Barrios, J. M. Saavedra, F. Ramirez, and D. Contreras. Orand at trecvid 2014: Instance search and multimedia event detection. In *Proc. of TRECVID*. NIST, USA, 2014.

[8] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[9] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *Proc. of the int. conf. on Computer Vision Theory and Application (VISSAPP)*, pages 331–340. INSTICC Press, 2009.

[10] P. Over, G. Awad, J. Fiscus, M. Michel, D. Joy, A. F. Smeaton, W. Kraaij, G. Quenot, and R. Ordelman. Trecvid 2015 – an overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2015*. NIST, USA, 2015.

[11] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. of the int. workshop on Multimedia Information Retrieval (MIR)*, pages 321–330. ACM, 2006.

[12] T. Tommasi, R. B. N. Aly, K. McGuinness, K. Chatfield, R. Arandjelovic, O. Parkhi, R. J. F. Ordelman, A. Zisserman, and T. Tuytelaars. Beyond metadata: searching your archive based on its audio-visual content. In *Proceedings of the 2014 International Broadcasting Convention, IBC 2014, Amsterdam, The Netherlands*, Stevenage, Herts, UK, September 2014. Institution of Engineering and Technology (IET).