HAL
open science

# Mining visual actions from movies

Adrien Gaidon, Marcin Marszalek, Cordelia Schmid

## To cite this version:

Adrien Gaidon, Marcin Marszalek, Cordelia Schmid. Mining visual actions from movies. British Machine Vision Conference, British Machine Vision Association, Sep 2009, Londres, United Kingdom. pp.128. inria-00440973v1

**HAL Id: inria-00440973**

**https://inria.hal.science/inria-00440973v1**

Submitted on 14 Dec 2009 (v1), last revised 25 Apr 2012 (v2)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Mining visual actions from movies

Adrien Gaidon[1]
http://lear.inrialpes.fr/people/gaidon/
Marcin Marszałek[2]
http://www.robots.ox.ac.uk/~marcin/
Cordelia Schmid[1]
http://lear.inrialpes.fr/people/schmid/

[1] LEAR
INRIA, LJK
Grenoble, France
[2] Visual Geometry Group
University of Oxford
Oxford, UK

Key frames of the top 5 "walk" and "kiss" samples obtained with our *iter-SVR* method (FP is the first false positive - "walk": rank 30, "kiss": 37).

This paper presents an approach for mining visual actions from real-world videos, using both text and vision. Following recent advances on action recognition [1, 4], we use movies and their transcripts to automatically obtain video samples of visual actions. Such a system can visually discover which actions are performed and also permits to collect training data for action recognition. Similar approaches were explored to build collections of images for object classes [7] and for naming characters in images [5] and videos [3].

**Overview.** First, we find commonly occurring actions by **mining verbs** extracted from movie transcripts. Next, we align the transcripts with the videos using subtitles. We then retrieve video samples for each action of interest. In practice, such a retrieval process includes visually irrelevant results. Therefore, we propose to rank the retrieval results using **visual consistency**. Following [4], our visual representation of videos is based on **bags of spatio-temporal visual words**, *i.e.* histograms of quantized local features. Our features are HOG descriptors computed from small 3D cuboids around multi-scale spatio-temporal Harris interest points We use a $\chi^2$ kernel to measure the similarity between bag-of-features representations. To quantify the visual consistency of a video clip in a retrieved list, we first consider two unsupervised outlier detection methods. We then show how to use weak supervision for ranking.

**Unsupervised estimation of inconsistency.** We first investigate **one-class Support Vector Machines** (SVM) [6]. The main idea is to find a hyper-sphere, in a high-dimensional space induced by a kernel, around the main consistent part of the data. We consider the distance from this boundary of normality, *i.e.* the distance from the margin obtained by the SVM, as an inconsistency score. The other unsupervised outlier removal technique we use is based on **densest components** of similarity graphs [5]. We first represent visual similarities between the retrieved samples as a graph structure. Then, we find the densest sub-graph by iteratively deleting the nodes with minimal degree and updating the graph. We rank our video samples by inverse pruning order. This defines the most inconsistent samples as the ones lying in the sparsest regions.

**Ranking with weak supervision.** As an alternative, we propose to rank clips using weak supervision provided by automatic annotations. In our setup, the subset of documents we want to rank is considered as positives, and randomly sampling the rest of the video collection is an inexpensive way to model inconsistent samples (considered as negatives). We show how to use such weak supervision to obtain a ranking function reflecting visual consistency. We first investigate the efficiency of a **binary $\nu$-SVM** [8]. It is a reformulation of the standard SVM optimization problem, replacing the regularization parameter $C$ with $\nu$, an upper bound on the fraction of outliers. Similarly to the one-class SVM case, we use a $\chi^2$ kernel and the distance from the margin, separating consistent samples from inconsistent ones, as a measure of visual inconsistency. However, the fact that we use weak supervision is not explicitly taken into account. Therefore, we also propose an iterative re-training heuristic for Support Vector Regression machines (SVR) [2], referred to as **"iter-SVR"**. We reformulate the ranking problem as a regression one. Our goal is to learn, from the weakly annotated data, a function that gives high values to consistent samples. The first step of our *iter-SVR* algorithm is to assign the target values $+1$ to all the retrieved samples, and $-1$ to all the random negative samples. We then train a SVR on this data and compute the obtained regressed values of the retrieved samples. We consider the normalized regression outputs of the SVR as new target values for our retrieved samples. We then re-train the SVR with this new improved supervision and repeat these steps until convergence.

**Experiments.** To evaluate our approach, we choose the TV show *Buffy the vampire slayer* in order to perform large-scale action retrieval. We evaluate our ranking algorithms on six manually selected action classes, 'walk', 'fall', 'punch', 'kick', 'kiss' and 'get-up'. Figure 1 illustrates the improvement due to the iterative re-labeling and re-ranking process performed by the "iter-SVR" algorithm. It improves the text-based retrieval results by $+12.4\%$ in mean Average Precision. We also demonstrate that weakly supervised algorithms are more efficient than unsupervised methods. Finally, we show that our *iter-SVR* algorithm can efficiently benefit from a weak supervision and outperforms the other commonly used approaches.

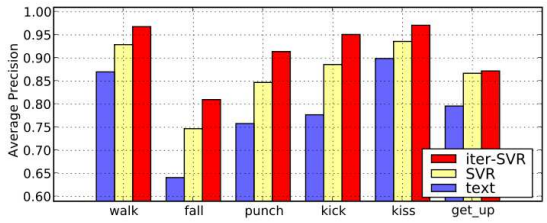


Figure 1: Comparison of our iterative SVR approach ("iter-SVR"), the SVR without re-training ("SVR") and the text mining results for six action classes.

[1] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/Script: Alignment and Parsing of Video and Text Transcription. In *ECCV*, 2009.
[2] H. Drucker, C.J.C. Burges, L. Kaufman, A. Smola, and V. Vapnik. Support vector regression machines. In *NIPS*, 1997.
[3] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy – Automatic Naming of Characters in TV Video. In *BMVC*, 2006.
[4] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.
[5] D. Ozkan and P. Duygulu. A graph based approach for naming faces in news photos. In *CVPR*, 2006.
[6] B. Schölkopf, R. Williamson, A. Smola, and J. Shawe-Taylor. SV estimation of a distribution's support. In *NIPS*, 1999.
[7] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.
[8] V.N. Vapnik. *The Nature of Statistical Learning Theory*. Springer, 2000.