Modeling Video Evolution For Action Recognition

Basura Fernando, Efstratios Gavves, José Oramas M., Amir Ghodrati, Tinne Tuytelaars
KU Leuven, ESAT, PSI, iMinds, Leuven, Belgium.

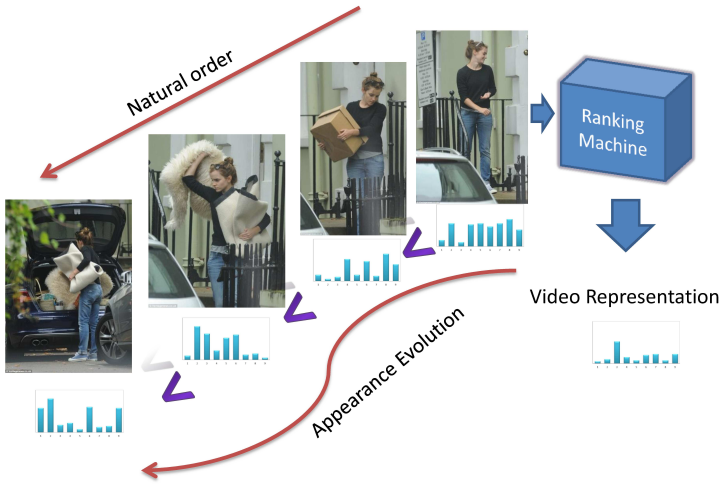

Figure 1: Illustration of how *VideoDarwin* works. In this video, as Emma *moved out from the house*, the appearance of the frames evolves with time. A ranking machine learns this evolution of the appearance over time and returns a ranking function. We use the parameters of this ranking function as a new video representation which captures vital information about the action.

Many actions have a characteristic temporal ordering - see e.g. the "moving out of the house" action in Figure 1. Intuitively, one would expect that a video representation that encodes this temporal ordering should help to better distinguish between different actions. As videos consist of sequences of images, obtaining a good video-wide representation remains a challenge.

In this paper, we approach a new video representation that captures this video-wide temporal evolution. We start from the observation that, even if the execution time of actions varies greatly, the *temporal ordering is typically preserved*. We propose to capture the temporal ordering *of a particular video* by training a linear ranking machine on the frames of that video. More precisely, given all the frames of the video, we learn how to arrange them in chronological order, based on the content of the frames. The parameters of the linear ranking functions encode the video-wide temporal evolution of appearance of videos in a principled way. To learn such ranking machines, we use the supervised learning to rank framework. Ranking machines trained on different videos of the same action can be expected to have similar ranking functions. Therefore, we propose to use the parameters of the ranking machine as a new video representation for action recognition. Classifiers trained on this new representation turn out to be remarkably good at distinguishing actions. Since the ranking machines act on frame content (in our experiments local spatio-temporal descriptors), they actually capture both the appearance and their evolution over time. We call our method *VideoDarwin* .

Our key contribution is to use the parameters of the ranking functions as a new video representation that captures the *video-wide temporal evolution of the video*. Our new video representation is based on a principled learning approach, it is easy to implement and efficient. Last but not least, with the new representation we obtain state-of-the art results in action and gesture recognition.

We start from a video $X = [\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n}]$ composed of n frames and frame at t is represented by vector $\mathbf{x_t} \in \mathbb{R}^D$. Each video X_i has a different evolution of appearances over time and will therefore learn a different ranking function $\psi_i = \psi(\cdot; \mathbf{u_i})$. As the ranking function ψ_i is video specific, we propose to use the parameters $\mathbf{u_i} \in \mathbb{R}^D$ of ψ_i as a new video representation to capture the specific *appearance evolution of the video*. Thus we obtain a functional representation, where the functional parameters $\mathbf{u_i}$ of the ranking function ψ_i serve as a representation that captures a vital part of the video-wide temporal information. We refer to our method as *VideoDarwin* since it exploits the evolution of appearance of a video.

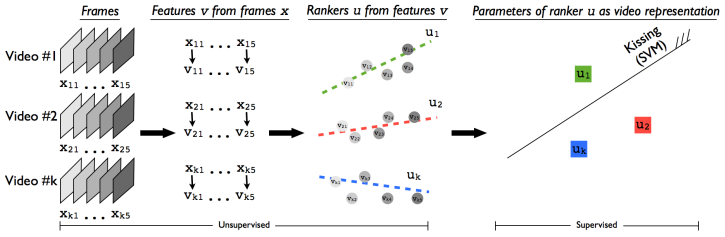

Figure 2: Processing steps of *VideoDarwin* for action recognition. First, we extract frames $\mathbf{x_1} \ldots \mathbf{x_n}$ from each video. Then we generate feature $\mathbf{v_t}$ for frame t by processing frames from $\mathbf{x_1}$ to $\mathbf{x_t}$. Afterwards, using ranking machines we learn the video representation $\mathbf{u}$ for each video. Finally, video specific $\mathbf{u}$ vectors are used as a representation for action classification.

	HMDB51	Hollywood2	Cooking
VideoDarwin	63.7	**73.7**	**72.0**
Previous Best	**66.8** [3]	73.6 [1]	70.5 [5]

Table 1: Comparison of the proposed approach with the state-of-the-art methods sorted by reverse chronological order. Results reported in mAP for Hollywood2 and Cooking datasets. For HMDB51 we report one-vs-all classification accuracy.

Although the optimization objective can be expressed on the basis of for example the RankSVM [2], any other linear learning to rank method can be employed to learn *VideoDarwin* . We incorporate non-linear families of functions by non-linear feature maps [4] applied on frame data. For action recognition we make the basic assumption that similar actions in different videos will have similar video wide temporal information. We use ranking machines that has certain stability guarantees to make sure this assumption is satisfied.

VideoDarwin is an unsupervised, learning based temporal pooling method, which aggregates the relevant information throughout a video via a learning to rank methodology. The generalization properties, as well as the regularized learning, of the learning to rank algorithms allow us to capture the global temporal information of a video while minimizing the empirical risk. As a result we arrive at a robust video representation suitable for action recognition. Based on extensive experimental evaluations on different datasets and features we conclude that, our method is applicable to any frame based representations for capturing the global temporal information of a video (see Table 1).

[1] Minh Hoai and Andrew Zisserman. Improving human action recognition using score distribution and ranking. In *ACCV*, 2014.

[2] Thorsten Joachims. Training linear svms in linear time. In *ICKDD*, 2006.

[3] X. Peng, C. Zou, Y. Qiao, and Q. Peng. Action recognition with stacked fisher vectors. In *ECCV*, 2014.

[4] Andrea Vedaldi and Andrew Zisserman. Efficient additive kernels via explicit feature maps. *PAMI*, 34:480–492, 2012.

[5] Y. Zhou, B. Ni, S. Yan, P. Moulin, and Q. Tian. Pipelining localized semantic features for fine-grained action recognition. In *ECCV*, 2014.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.