Actions in Context

Marcin Marszałek
INRIA Grenoble
marcin.marszalek@inria.fr

Ivan Laptev
INRIA Rennes
ivan.laptev@inria.fr

Cordelia Schmid
INRIA Grenoble
cordelia.schmid@inria.fr

Abstract

This paper exploits the context of natural dynamic scenes for human action recognition in video. Human actions are frequently constrained by the purpose and the physical properties of scenes and demonstrate high correlation with particular scene classes. For example, eating often happens in a kitchen while running is more common outdoors. The contribution of this paper is three-fold: (a) we automatically discover relevant scene classes and their correlation with human actions, (b) we show how to learn selected scene classes from video without manual supervision and (c) we develop a joint framework for action and scene recognition and demonstrate improved recognition of both in natural video. We use movie scripts as a means of automatic supervision for training. For selected action classes we identify correlated scene classes in text and then retrieve video samples of actions and scenes for training using script-to-video alignment. Our visual models for scenes and actions are formulated within the bag-of-features framework and are combined in a joint scene-action SVM-based classifier. We report experimental results and validate the method on a new large dataset with twelve action classes and ten scene classes acquired from 69 movies.

1. Introduction

Video becomes an easily created and widely spread media serving entertainment, education, communication and other purposes. The associated demand for mining large collections of realistic video data motivates further research in automatic video understanding.

Video understanding involves interpretation of objects, scenes and actions. Whereas previous work mostly dealt with these concepts separately, a unified approach is expected to provide yet unexplored opportunities to benefit from mutual contextual constraints among actions, scenes and objects. For example, while some actions and scenes typically co-occur, the chance of co-occurrence of random classes of scenes and actions can be low.

In this work, we build upon the above intuition and exploit co-occurrence relations between actions and scenes in video. Starting from a given set of action classes, we aim to automatically discover correlated scene classes and to use this correlation to improve action recognition. Since some actions are relatively scene-independent (e.g. "smile"), we do not expect context to be equally important for all actions. Scene context, however, is correlated with many action classes. Moreover, it often defines actions such as "lock the door" or "start the car". It is therefore essential for action recognition in general.

(a) eating, kitchen (b) eating, cafe

(c) running, road (d) running, street

Figure 1. Video samples from our dataset with high co-occurrences of actions and scenes and automatically assigned annotations.

We aim to explore scenes and actions in generic and realistic video settings. We avoid specific scenarios such as surveillance or sports, and consider a large and diverse set of video samples from movies, as illustrated in Fig. 1. We use movie scripts for automatic video annotation and apply text mining to discover scene classes which co-occur with given actions. We use script-to-video alignment and text search to automatically retrieve video samples and corresponding labels for scenes and actions in movies. Note, that we only use scripts for training and do not assume scripts to be available during testing.

Given the automatically retrieved video samples with possibly noisy labels, we use the bag-of-features representation and SVM to learn separate visual models for action and scene classification. As the main contribution of this paper we demonstrate that automatically estimated contextual relations improve both (i) recognition of actions in the context of scenes as well as (ii) recognition of scenes in the

978-1-4244-3991-1/09/$25.00 ©2009 IEEE

context of actions. We, in particular, emphasize the fully automatic nature of our approach and its scalability to a large number of action and scene classes.

The rest of the paper is organized as follows. The remaining part of this section reviews related work. Section 2 presents script mining for discovering contextual relations and automatic video annotation. Section 3 describes independent visual classification of actions and scenes. Section 4 introduces the joint action and scene recognition and evaluates the contextual relationships mined from text and learned from visual data. Section 5 concludes the paper.

1.1. Related work

Visual context has a long history of research with early studies in cognitive science [1]. In computer vision context has been used for interpretation of static scenes [10, 20, 21, 26]. For example, scene context is used to impose spatial priors on the location of objects in the image [21, 26]. In a similar spirit, [20] exploits spatial relations between objects and scene parts for segmentation and recognition of static images. Co-occurrence of people in images is explored for face recognition in [16].

A few papers explore the context of human actions. Similar to our work, [14] exploits scene context for event recognition, but only applies it to static images. In [8, 18, 22, 27] object context is used for action recognition and demonstrates improved recognition of objects and actions in video. However, in most of the previous work only constrained experimental settings are considered. In contrast, we focus on action recognition in realistic video data from movies. Furthermore, we automatically discover contextual relations between scenes and actions.

Scripts have been explored as a means of automatic video annotation in [2, 5, 12]. We follow this work and add the use of scripts to estimate relations between scenes and actions in video. A related approach deriving object relations from textual annotations in still images has been recently presented in [9]. For action and scene recognition we follow bag-of-features models explored in many recent works on object, scene and action recognition in still images and video [3, 4, 12, 13, 19, 24, 25].

2. Script mining for visual learning

Learning visual representations of realistic human actions and scenes requires a large amount of annotated video data for training. While manual video annotation and data collection is time-consuming and often unsatisfying, recent work adopts video-aligned scripts for automatic annotation of videos such as movies and sitcoms [2, 5, 12]. In the following we briefly describe script-to-video alignment in subsection 2.1 and script-based retrieval of action samples in subsection 2.2. Subsection 2.3 presents automatic script mining of action-correlated scene classes and discusses the retrieval of the corresponding video samples.

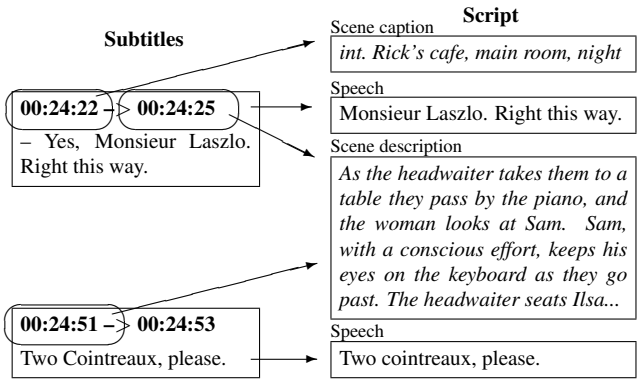


Figure 2. Script synchronization using timestamps from subtitles.

2.1. Script-to-video alignment

Scripts are text documents publicly available for the majority of popular movies. They contain scene captions, dialogs and scene descriptions, but usually do not provide time synchronization with the video. Following [2, 5, 12], we address this problem by synchronizing script dialogs with the corresponding subtitles. Subtitles are easy to obtain from the web or DVDs and are already synchronized with the video through timestamps. Hence, we match script text with subtitles using dynamic programming and then estimate temporal localization of scene captions and scene descriptions by transferring time information from subtitles to scripts. Using this procedure, illustrated in Fig. 2, we obtain a set of short video clips (segmented by subtitle timestamps) with corresponding script parts, i.e., textual descriptions.

2.2. Action retrieval

To automatically collect video samples for human actions, we follow the approach in [12]. We choose twelve frequent action classes and manually assign corresponding action labels to scene descriptions in a few movie scripts. Note that these scripts are distinct from the 69 scripts used in the following for automatic annotation. We then train an off-the-shelf bag-of-words text classifier and automatically retrieve scene descriptions with action labels from a set of 69 movie scripts[1] synchronized with the video as described

[1]We obtain movie scripts from www.dailyscript.com, www.movie-page.com and www.weeklyscript.com. The 69 movies used in this paper are divided into 33 training movies and 36 test movies as follows.

Training movies: American Beauty, As Good as It Gets, Being John Malkovich, The Big Lebowski, Bruce Almighty The Butterfly Effect, Capote, Casablanca, Charade, Chasing Amy, The Cider House Rules, Clerks, Crash, Double Indemnity, Forrest Gump, The Godfather, The Graduate, The Hudsucker Proxy, Jackie Brown, Jay and Silent Bob Strike Back, Kids, Legally Blonde, Light Sleeper, Little Miss Sunshine, Living in Oblivion, Lone Star, Men in Black, The Naked City, Pirates of the

	Auto Train	Clean Test
AnswerPhone	59	64
DriveCar	90	102
Eat	44	33
FightPerson	33	70
GetOutCar	40	57
HandShake	38	45
HugPerson	27	66
Kiss	125	103
Run	187	141
SitDown	87	108
SitUp	26	37
StandUp	133	146
All Samples	810	884

(a) Actions

	Auto Train	Clean Test
EXT-House	81	140
EXT-Road	81	114
INT-Bedroom	67	69
INT-Car	44	68
INT-Hotel	59	37
INT-Kitchen	38	24
INT-LivingRoom	30	51
INT-Office	114	110
INT-Restaurant	44	36
INT-Shop	47	28
All Samples	570	582

(b) Scenes

Table 1. Distribution of video samples in two automatically generated training sets and two manually verified test sets for action and scene classes respectively. The total length of action samples is about 600k frames or 7 hours of video. For scene samples it is about 990k frames or 11 hours. Our dataset is publicly available at `http://www.irisa.fr/vista/actions/hollywood2`

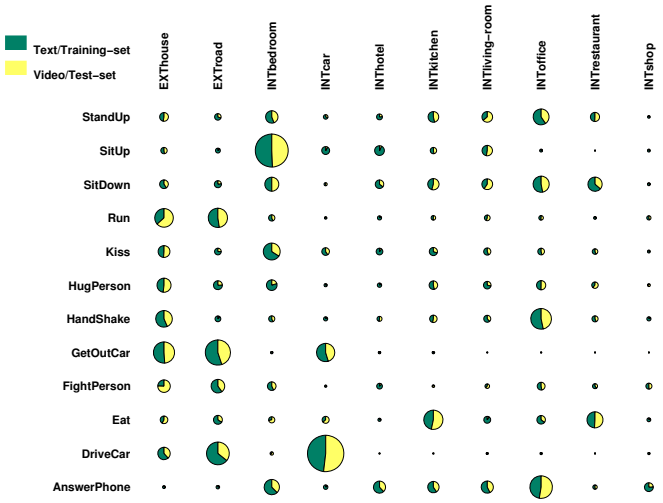


Figure 3. Conditional probabilities $p(Scene|Action)$ estimated from scripts (green) and ground truth visual annotation (yellow). Note the consistency of high probability values (large circles) with intuitively expected correlations between actions and scenes. Observe also the consistency of probability values automatically estimated from text and manually estimated from visual video data.

in Section 2.1. With this procedure we automatically obtain video samples for a chosen set of action labels.

We split all retrieved samples into the test and training subsets, such that the two subsets do not share samples from the same movie. For training samples we keep the (noisy) labels obtained with the automatic procedure. We manually validate and correct labels for the test samples in order to properly evaluate the performance of our approach. Table 1 (left) summarizes the action classes and the numbers of samples we use for training and testing action classifiers.

2.3. Scene retrieval

The main objective of this work is to exploit re-occurring relations between actions and their scene context to improve classification. For a given set of action classes, this requires both (i) identification of relevant scene types and (ii) estimation of the co-occurrence relation with actions. We aim to solve both tasks automatically in a scalable framework. While visual data could be used for our purpose, we find it easier to discover relevant scene types in text. As in the previous section, we resort to movie scripts. We use *scene captions*, short descriptions of the scene setup, which are consistently annotated in most of the movie scripts and usually provide information on location and day time:

INT. TRENDY RESTAURANT - NIGHT
INT. M. WALLACE'S DINING ROOM - MORNING
EXT. STREETS BY DORA'S HOUSE - DAY.

To discover relevant scene concepts, we collect unique words and consecutive word pairs from the captions. We use WordNet [7] to select expressions corresponding to instances of "physical entity". We also use the WordNet hierarchy to generalize concepts to their hyponyms, such as *taxi→car*, *cafe→restaurant*, but preserve concepts which share hyponyms, such as *kitchen*, *living room* and *bedroom* (they share the *room* hyponym). We also explicitly recognize INT (interior) and EXT (exterior) tags.

From the resulting 2000 contextual concepts we select 30 that maximize co-occurrence with action classes in the training scripts. To select both frequent and informative concepts, we sort them by the entropy computed for the distributions of action labels. This results in an ordered list from which we take the top ten scene concepts.

Figure 3 illustrates co-occurrences between automatically selected scene classes and actions. The size of the circles corresponds to the estimated probabilities of scenes for given action classes and coincides well with intuitive expectations. For example, *running* mostly occurs in outdoor scenes while *eating* in kitchen and restaurant scenes.

Caribbean: Dead Man's Chest, Psycho, Quills, Rear Window, Fight Club.

Test movies: Big Fish, Bringing Out The Dead, The Crying Game, Dead Poets Society, Erin Brockovich, Fantastic Four, Fargo, Fear and Loathing in Las Vegas, Five Easy Pieces, Gandhi, Gang Related, Get Shorty, The Grapes of Wrath, The Hustler, I Am Sam, Independence Day, Indiana Jones and The Last Crusade, It Happened One Night, It's a Wonderful Life, LA Confidential, The Lord of the Rings: The Fellowship of the Ring, Lost Highway, The Lost Weekend, Midnight Run, Misery, Mission to Mars, Moonstruck, Mumford, The Night of the Hunter, Ninotchka, O Brother Where Art Thou, The Pianist, The Princess Bride, Pulp Fiction, Raising Arizona, Reservoir Dogs.

Since the selection of scene classes and their relations to actions were automatically derived from the training movie scripts, we validate (i) how well script-estimated co-occurrences transfer to the video domain and (ii) how well they generalize to the test set. We manually annotate scenes in test videos and perform evaluation. Co-occurrences estimated automatically from training scripts and manually evaluated on test movies are illustrated with different colors in Fig. 3. An equal division of the circles by colors validates the consistency between relations mined from text and those actually occurring in videos.

Having found correlated scene concepts, we automatically retrieve the corresponding video samples as in Section 2.2. Note, however, that there is no need for a text classifier in this case. The procedure for finding correlated scene concepts is therefore fully automatic and unsupervised. As for the actions dataset, we keep the automatically obtained scene labels for the training set and verify the labels to be correct for the test set. The automatically selected scene classes and the number of corresponding samples in both subsets are illustrated in Table 1 (right). Note that for training *we do not use any manual annotation* of video samples neither for action nor scene classes.

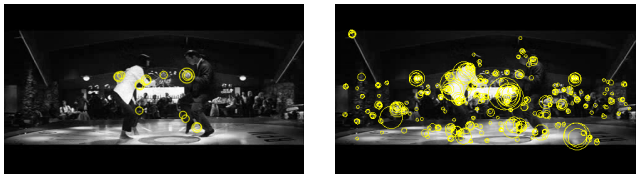
3. Visual learning of actions and scenes

Our actions and scene classifiers are based on a bag-of-features image representation [3, 4, 25] and classification with Support Vector Machines [23]. Our video representation uses simultaneously static and dynamic features as described in subsection 3.1. In subsection 3.2 we briefly introduce the approach for classification and subsection 3.3 evaluates features for action and scene classification.

3.1. Bag-of-features video representation

Inspired by the recent progress of recognition in unconstrained videos [12], we build on the bag-of-features video representation for action recognition [19, 24]. The bag-of-features representation views videos as spatio-temporal volumes. Given a video sample, salient local regions are extracted using an interest point detector. Next, the video content in each of the detected regions is described with local descriptors. Finally, orderless distributions of features are computed for the entire video sample.

To construct our baseline we follow the work of Laptev [11] and extract motion-based space-time features. This representation focuses on human actions viewed as motion patterns. The detected salient regions correspond to motion extrema and features are based on gradient dynamics and optical flow. This baseline approach has shown success for action recognition, but is alone not sufficient for scene classification. To reliably model scenes, we need to include static appearance in addition to motion patterns. We



(a) 3D Harris (b) 2D Harris

Figure 4. Interest point detections for a movie frame. Note that 3D Harris points (left) focus on motion, whereas 2D Harris points (right) are distributed over the scene.

therefore view the video as a collection of frames as well. This allows us to employ the bag-of-features components developed for static scene and object recognition [28] and have a hybrid representation in a unified framework. In the following we describe each component in detail.

3D-Harris detector. Salient motion is detected using a multi-scale 3D generalization of the *Harris* operator [11]. Since the operator responds to 3D corners, applying it to video volumes allows to detect (in space and time) characteristic points of moving salient parts. The left image of figure 4 shows an example where the detector responds to actors dancing in the center of the scene.

2D-Harris detector. To describe static content we use the 2D scale-invariant Harris operator [17]. It responds to corner-like structures in images and allows us to detect salient areas in individual frames extracted from the video. See the right part of Fig. 4 for an example, where the detector responds to salient areas all around the scene. We detect static features in a video stream every second.

HoF and HoG descriptors. We compute *HoG* and *HoF* descriptors [12] for the regions obtained with the 3D-Harris detector. HoF is based on local histograms of optical flow. It describes the motion in a local region. HoG is a 3D histogram of 2D (spatial) gradient orientations. It describes the static appearance over space and time.

SIFT descriptor. We compute the *SIFT* descriptor [15] for the regions obtained with the 2D-Harris detector. SIFT is a weighted histogram of gradient orientations. Unlike the HoG descriptor, it is a 2D (spatial) histogram and does not take into account the temporal domain.

These three different types of descriptors capture motion patterns (HoF), dynamic appearance (HoG) and static appearance (SIFT). For each descriptor type we create a visual vocabulary [25]. These vocabularies are used to represent a video sample as an orderless distribution of visual words, typically called *bag-of-features* [3]. We compare feature distributions using χ^2 distance and fuse different feature types in the classifier as described in the following.

3.2. χ^2 Support Vector Classifier

We use Support Vector Machines [23] (SVM) to learn actions and scene context. The decision function of a binary

C-SVM classifier takes the following form:

$$g(\boldsymbol{x}) = \sum_i \alpha_i y_i K(\boldsymbol{x_i}, \boldsymbol{x}) - b \qquad (1)$$

where $K(\boldsymbol{x_i}, \boldsymbol{x})$ is the value of a kernel function for the training sample $\boldsymbol{x_i}$ and the test sample $\boldsymbol{x}$, $y_i \in \{+1, -1\}$ is the class label (positive/negative) of $\boldsymbol{x_i}$, α_i is a learned weight of the training sample $\boldsymbol{x_i}$, and b is a learned threshold. We use the values of the decision function as a confidence score and plot recall-precision curves for evaluation. We then compute the average precision (AP), which approximates the area under a recall-precision curve [6].

The simplest kernel possible is a linear kernel. The decision function can then be rewritten as a weighted sum of sample components, i.e.,

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \boldsymbol{x_i} \cdot \boldsymbol{x_j} \ , \qquad g(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} - b \ . \qquad (2)$$

To classify feature distributions compared with the χ^2 distance, we use the multi-channel Gaussian kernel:

$$K(\boldsymbol{x_i}, \boldsymbol{x_j}) = \exp\big(-\sum_c \frac{1}{\Omega_c} D_c(\boldsymbol{x_i}, \boldsymbol{x_j})\big) \qquad (3)$$

where $D_c(\boldsymbol{x_i}, \boldsymbol{x_j})$ is the χ^2 distance for channel c, and Ω_c is the average channel distance [28].

To build a multi-class classifier one might combine binary classifiers using *one-against-rest* or *one-against-one* strategies. Note, however, that in our setup all problems are binary, i.e., we recognize each concept independently and concurrent presence of multiple concepts (mainly multiple actions) is possible. To compare the overall system performance, we compute an average AP (AAP) over a set of binary classification problems.

3.3. Evaluation of the action and scene classifiers

Figure 5 compares HoG, HoF and SIFT features for action and scene recognition on the datasets described in Section 2. On the left of Fig. 5 we show classes where SIFT features give better results than the other feature types. The majority of these classes are scenes types, both interior and exterior. It is interesting to observe that there are also the actions *sitting up* and *getting out of car*. This can be explained by the context information provided by the bedroom setting and the presence of a car respectively. On the right side of Fig. 5 we display classes where HoG and HoF features are more appropriate. They include actions like *fighting*, *sitting down* and *standing up*. Interestingly, we also observe action and scene classes related to driving a car. This could be due to the characteristic background motion seen through the vehicle windows while driving.

Overall, the results confirm that both static and dynamic features are important for recognition of scenes and actions. Furthermore, the performance of a feature type might be difficult to predict due to the influence of the context.

To exploit the properties of different feature types, we combine them. Table 2 compares static SIFT features, the combination of dynamic HoGs and HoFs, and the combination of all three feature types. Note that the combination of HoGs and HoFs corresponds to the state-of-the-art action recognition method of Laptev *et al*. [12].

The two leftmost columns of Table 2 quantify and allow further analysis of previously discussed results (cf. Fig. 5). The comparison confirms that some action classes are better recognized with static features, whereas some scene classes can be well characterized with dynamic features. Yet, static features are generally better for scene recognition (7/10 classes, in bold) and dynamic features for action recognition (8/12 classes, in bold). This result also confirms our hypothesis that the baseline video representation using only dynamic features is not suitable for scene recognition.

The rightmost column of Table 2 shows the recognition performance when all three feature types are combined. For action classes (upper part), combining static SIFT features with dynamic HoG and HoF features improves accuracy in 8/12 cases (in bold). Yet, the average action classification

	SIFT	HoG HoF	SIFT HoG HoF
AnswerPhone	**0.105**	0.088	**0.107**
DriveCar	0.313	**0.749**	0.750
Eat	0.082	**0.263**	**0.286**
FightPerson	0.081	**0.675**	0.571
GetOutCar	**0.191**	0.090	**0.116**
HandShake	**0.123**	0.116	**0.141**
HugPerson	0.129	**0.135**	**0.138**
Kiss	0.348	**0.496**	**0.556**
Run	0.458	**0.537**	**0.565**
SitDown	0.161	**0.316**	0.278
SitUp	**0.142**	0.072	**0.078**
StandUp	0.262	**0.350**	0.325
Action average	*0.200*	*0.324*	*0.326*
EXT.House	**0.503**	0.363	0.491
EXT.Road	**0.498**	0.372	0.389
INT.Bedroom	**0.445**	0.362	**0.462**
INT.Car	0.444	**0.759**	**0.773**
INT.Hotel	0.141	**0.220**	**0.250**
INT.Kitchen	**0.081**	0.050	0.070
INT.LivingRoom	0.109	**0.128**	**0.152**
INT.Office	**0.602**	0.453	0.574
INT.Restaurant	**0.112**	0.103	0.108
INT.Shop	**0.257**	0.149	0.244
Scene average	*0.319*	*0.296*	*0.351*
Total average	*0.259*	*0.310*	*0.339*

Table 2. Comparison of feature combinations in our bag-of-features approach. Average precision is given for action and scene classification in movies. Both actions and scenes benefit from combining static and dynamic features.

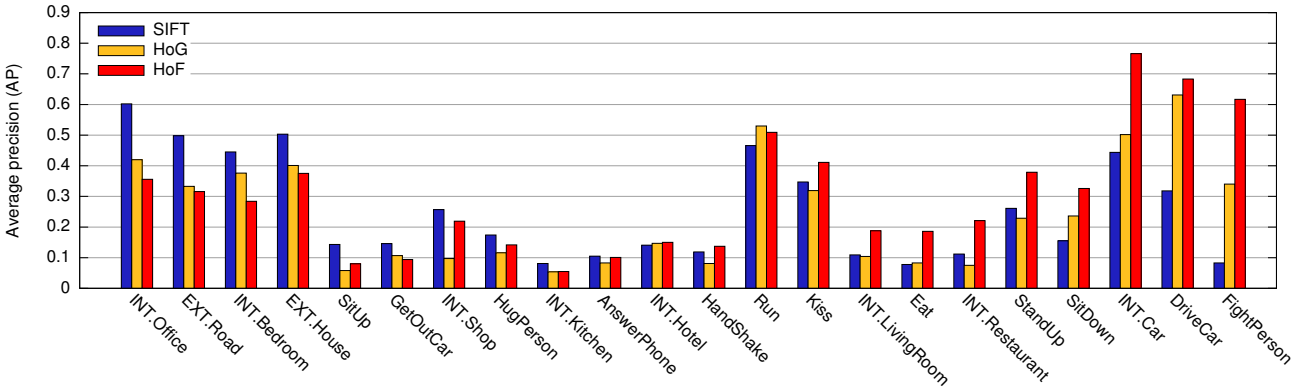


Figure 5. Comparison of single feature types in our bag-of-features approach. Average precision is given for action and scene classification in movies. The static SIFT features perform well for most scene types while HoG and HoF features are important for action recognition.

score remains basically the same as for the HoG and HoF baseline. This is due to the loss in classes like *fighting*, *sitting down* and *standing up* which are strongly motion-based and where static features might only introduce unnecessary background noise. For scene classes (lower part), a combination of all feature types improves over SIFT for only 4/10 cases (in bold). Yet, the improvement for scene classification is significant on average. Combining static and dynamic features sometimes degrades performance. However, for classes like *car interior*, where motion plays a critical role, it results in a significant gain.

Summary. The results show that a combination of static and dynamic features is a good choice for a generic video recognition framework. It allows to recognize actions and scenes in a uniform setup. Our approach also models implicit context, i.e., events which are not directly related to the action, such as background motion in *car interior*.

4. Classification with context

The goal in this section is to improve action classification based on scene context obtained with the scene classifiers. We integrate context by updating the classification score $g_a(\boldsymbol{x})$, cf. (1), for an action $a \in \mathcal{A}$. To the original classification score we add a linear combination of the scores $g_s(\boldsymbol{x})$ for contextual concepts (scenes) $s \in \mathcal{S}$:

$$g'_a(\boldsymbol{x}) = g_a(\boldsymbol{x}) + \tau \sum_{s \in \mathcal{S}} w_{as} g_s(\boldsymbol{x}) \qquad (4)$$

where τ is a global context weight and w_{as} are weights linking concepts a and s.

Given a set $\mathcal{A}$ of actions and a set $\mathcal{S}$ of contextual categories for which we have g_a and g_s decision functions respectively, the parameters of our context model are τ and w_{as}. In this paper we propose two methods for obtaining the weights w_{as}. The first one is based on text mining[2] and the second one learns the weights from visual data. The τ parameter allows to control the influence of the context. We have observed that the results are not very sensitive to this parameter and set it to a value of 3 for all our experiments. Classification is always based on the combination of all three feature types.

4.1. Mining contextual links from text

The most straightforward way to obtain the weights w_{as} is based on the information contained in the scripts, i.e., $p(Scene|Action)$ estimated from scripts of the training set as described in Section 2.3. As shown in Figure 3, these weights correspond well to our intuition and the visual information in videos.

Let us rewrite the context model given by (4) as

$$\boldsymbol{g'_a}(\boldsymbol{x}) = \boldsymbol{g_a}(\boldsymbol{x}) + \tau\, W \boldsymbol{g_s}(\boldsymbol{x}) \qquad (5)$$

where $\boldsymbol{g'_a}(\boldsymbol{x})$ is the vector of new scores and $\boldsymbol{g_a}(\boldsymbol{x})$ is a vector with the original scores for all basic (action) classes, $\boldsymbol{g_s}(\boldsymbol{x})$ is a vector with scores for all context (scene) classes, and W is a weight matrix.

We determine W from scripts of the training set by defining it to be a conditional probability matrix encoding the probability of a scene given an action $P(S \in \mathcal{S}|A \in \mathcal{A})$.

4.2. Learning contextual links from visual data

An alternative approach to determine contextual relationships between scenes and actions is to learn them from the visual data. We keep the same model as given by (4), but rewrite it as

$$g'_a(\boldsymbol{x}) = g_a(\boldsymbol{x}) + \tau\, \boldsymbol{w}_a \cdot \boldsymbol{g_s}(\boldsymbol{x}) \qquad (6)$$

where $\boldsymbol{w}_a$ is a vector of weights and $\boldsymbol{g_s}(\boldsymbol{x})$ is a vector formed with the scores for contextual concepts. We then

[2]Note that the weights are determined in the training stage and then kept for the recognition phase, so even with a text-mining approach no textual data is necessary for the recognition itself, only for training.

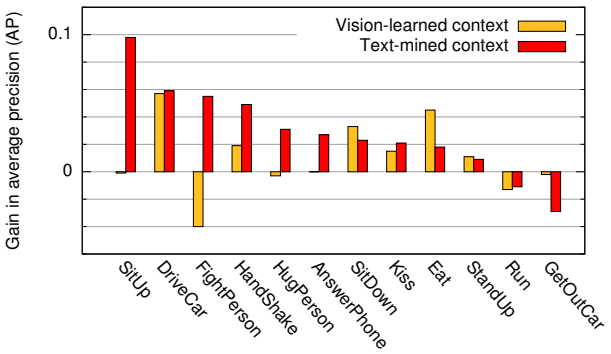


Figure 7. Exploiting scene context in action recognition. AP gain for SVM-learned and text-mined context is compared. Note the consistent improvement for most action classes.

replace the linear combination with a linear C-SVM classifier, cf. (2), denoted as z_a:

$$g'_a(\boldsymbol{x}) = g_a(\boldsymbol{x}) + \tau z_a(\boldsymbol{g_s}(\boldsymbol{x})) \ . \tag{7}$$

This allows us to learn the weights in a supervised manner, i.e., the weights of our context model are determined by a second-layer SVM classifier. This linear classifier is trained after the first-layer g_a and g_s classifiers are trained. It uses the same training data, but combines action and context classes in one learning task. Thus, from a perspective of a single action classifier, the information available for learning is richer. During recognition, the second-layer classifier takes a vector of scene scores as an input and produces a contextual prediction for the action class.

4.3. Experimental results

Action recognition. Figure 7 evaluates the gain for action recognition when using scene context. We show the average precision gains for context weights learned from text (red) and with a linear SVM (yellow). We obtain a gain of up to 10%, which is a remarkable improvement in difficult realistic setting. The largest gain is observed for action classes like *sitting up* or *driving a car*, which have a strong localized context of *bedroom* and *car* respectively, cf. Fig. 3. The more uniform scene context of *fighting* is easily exploited with text mining, but confuses the SVM classifier.

Context	AAP
text context	**0.355**
vision context	0.336
no context	0.325
context *only*	0.238
chance	0.125

(a) Actions

Context	AAP
text context	**0.373**
vision context	0.371
no context	0.351
context *only*	0.277
chance	0.162

(b) Scenes

Table 3. Average AP obtained using various context sources and not using context. We also compare to a model where context only is used and to chance level.

Figure 7 also shows that our method does not help much for action classes like *standing up*, which have low dependence on the scene type. Furthermore, the context mined from text leads to some confusion when trying to recognize *getting out of a car*. This might be due to special shooting conditions for this action class, which are typical for movies. Still, for most (75%) classes we observe a consistent and significant improvement of at least 2.5%, obtained with the context mined from scripts. This shows the difficulty of learning contextual relations using a limited amount of visual data, but at the same time highlights the ability of our method to overcome this problem by mining text.

Table 3 (left) compares average AP achieved with different context sources when classifying actions using scenes as context. It confirms that context weights improve the overall performance. The gain is at 3% for weights mined from text and at 2% for learned weights. Interestingly, using context *only*, i.e., classifying actions with only contextual scene classifiers, is still significantly better than chance. This shows the potential of context in natural videos.

Overall, the experiments confirm that our context model can exploit contextual relationships between scenes and actions. It integrates well with bag-of-features and SVM based classifiers and helps action recognition. See Fig. 6 for sample classification results where the scene context helped action recognition the most.

Scene recognition. We have observed that action classification can be improved by using contextual scene information. We will now investigate if the inverse also holds, i.e., whether we can improve scene classification by recognizing actions. Figure 8 evaluates the gain for scene recogni-

(a) DriveCar (b) FightPerson (c) HandShake (d) SitUp

Figure 6. Sample frames for action video clips where the context significantly helps recognition. Please note the *car interior* for *driving*, the *outdoor setting* for *fighting*, the *house exterior* for *handshaking*, and finally the *bedroom* for *sitting up*.

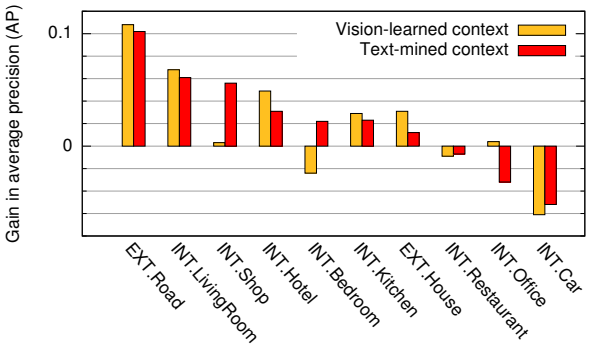


Figure 8. Exploiting action context in scene recognition. AP gain for SVM-learned and text-mined context is compared. Note the significant improvement for the leftmost categories.

tion when using action context. We show the average precision gains for context mined from textual data (red) and learned with a linear SVM (yellow). As for action recognition, cf. Fig. 7, an AP gain of up to 11% is observed. Again, an improvement of at least 2.5% is obtained for most of the classes. Concurrent analysis with Fig. 3 explains the difficulties of learning the context for *shop interior*. This scene type does not occur too often for any of the action classes. Table 3 (right) confirms these results, and for both types of context sources shows an improvement of 2% on average.

5. Conclusions

This paper shows that we can use automatically extracted context information based on video scripts and improve action recognition. Given completely automatic action and scene training sets we can learn individual classifiers which integrate static and dynamic information. Experimental results show that both types of information are important for actions and scenes and that in some cases the context information dominates. For example car interior is better described by motion features. The use of context information can be made explicit by modeling the relations between actions and scenes based on textual co-occurrence and by then combining the individual classifiers. This improves the results for action as well as scene classification. In this paper we have also created a new dataset which includes actions, scenes as well as their co-occurrence in real-world videos. Furthermore, we have demonstrated the usefulness of textual information contained in scripts for visual learning.

Acknowledgments. This work was partly funded by the Quaero project. We would like to thank B. Rozenfeld for his help with action classification in scripts.

References

[1] I. Biederman, R. Mezzanotte, and J. Rabinowitz. Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.*, 14:143–177, 1982.

[2] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription. In *ECCV*, 2008.

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision*, 2004.

[4] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[5] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... Buffy – automatic naming of characters in TV video. In *BMVC*, 2006.

[6] M. Everingham, L. van Gool, C. Williams, J. Winn, and A. Zisserman. Overview and results of the classification challenge, 2008. The PASCAL VOC'08 Challenge Workshop, in conj. with ECCV.

[7] C. Fellbaum, editor. *Wordnet: An Electronic Lexical Database*. Bradford Books, 1998.

[8] A. Gupta and L. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[9] A. Gupta and L. Davis. Beyond nouns: Exploiting prepositions and comparative adjectives for learning visual classifiers. In *ECCV*, 2008.

[10] G. Heitz and D. Koller. Learning spatial context: Using stuff to find things. In *ECCV*, 2008.

[11] I. Laptev. On space-time interest points. *IJCV*, 64(2/3):107–123, 2005.

[12] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008.

[13] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[14] L. Li and L. Fei-Fei. What, where and who? classifying events by scene and object recognition. In *ICCV*, 2007.

[15] D. Lowe. Distinctive image features form scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[16] T. Mensink and J. Verbeek. Improving people search using query expansions: How friends help to find people. In *ECCV*, 2008.

[17] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004.

[18] D. Moore, I. Essa, and M. Hayes. Exploiting human actions and object context for recognition tasks. In *ICCV*, 1999.

[19] J. C. Niebles, H. Wang, and L. Fei-Fei. Unsupervised learning of human action categories using spatial-temporal words. In *BMVC*, 2006.

[20] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *ICCV*, 2007.

[21] B. Russell, A. Torralba, C. Liu, R. Fergus, and W. Freeman. Object recognition by scene alignment. In *NIPS*, 2007.

[22] M. Ryoo and J. Aggarwal. Hierarchical recognition of human activities interacting with objects. In *CVPR*, 2007.

[23] B. Schölkopf and A. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

[24] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: A local SVM approach. In *ICPR*, 2004.

[25] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, volume 2, 2003.

[26] A. Torralba. Contextual priming for object detection. *IJCV*, 53(2):169–191, July 2003.

[27] J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. Rehg. A scalable approach to activity recognition based on object use. In *ICCV*, 2007.

[28] J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classification of texture and object categories: A comprehensive study. *IJCV*, 73(2):213–238, 2007.