Retrieving Actions in Group Contexts

Tian Lan[1], Yang Wang[1], Greg Mori[1], and Stephen N. Robinovitch[2]

[1] School of Computing Science, Simon Fraser University, Canada
[2] School of Kinesiology, Simon Fraser University, Canada

Abstract. We develop methods for action retrieval from surveillance video using contextual feature representations. The novelty of our proposed approach is two-fold. First, we introduce a new feature representation called the *action context (AC) descriptor*. The AC descriptor encodes information about not only the action of an individual person in the video, but also the behaviour of other people nearby. This feature representation is inspired by the fact that the context of what other people are doing provides very useful cues for recognizing the actions of each individual. Second, we formulate our problem as a retrieval/ranking task, which is different from previous work on action classification. We develop an action retrieval technique based on rank-SVM, a state-of-the-art approach for solving ranking problems. We apply our proposed approach on two real-world datasets. The first dataset consists of videos of multiple people performing several group activities. The second dataset consists of surveillance videos from a nursing home environment. Our experimental results show the advantage of using contextual information for disambiguating different actions and the benefit of using rank-SVMs instead of regular SVMs for video retrieval problems.

1 Introduction

In this paper we develop methods for human action retrieval from surveillance video data. Consider the video frames shown in Fig. 1. These are example frames from a nursing home surveillance video in which we would like to retrieve instances of actions of interest such as residents who fall. The intra-class variation in action categories and relatively poor video quality typical of surveillance footage render this a challenging problem. With this type of video footage many actions are ambiguous. For example, falling down and sitting down are often confused as shown in Fig. 1 – both can contain substantial downward motion and result in similarly shaped person silhouettes. A helpful cue that can be employed to disambiguate situations such as these is the context of what other people in the video are doing. Given visual cues of large downward motion, if we see other people coming to aid then it is more likely to be a fall than if we see other people sitting down.

We develop a novel representation to model this type of contextual interaction between the actions of individuals in a video. Our work employs a *bag-of-words* style representation, describing one person using his visual features along with the actions of others nearby. We demonstrate that this augmentation by

(a) (b)

Fig. 1. Role of context in action. It is often hard to distinguish actions from each individual person alone. An example is the action of falling down and sitting down performed by persons in the red bounding boxes in (a) and (b). However, if we look at what the people nearby (in the blue bounding boxes) are doing, the actions can be disambiguated.

including a representation of neighboring actions can improve action retrieval performance.

Bag-of-words representations have been studied extensively in computer vision, particularly in object recognition. In action recognition, Wang and Mori [1] track individual people and model co-occurrences of the actions in a single track with a mapping of frames to visual words. In contrast, the method we present here does not require tracking, which is challenging in our datasets, and models the actions of multiple people. Wang et al. [2] analyze far-field traffic video. Low-level atomic events are described by motion and position features, and hierarchical models are used to capture the co-occurances of these atomic events over video clips. We explicitly model the spatial context of an individual person, rather than treating the whole frame in a bag-of-words representation. Loy et al. [3] develop a structure learning algorithm to model temporal dependencies of actions across a camera network. Our model focuses on a lower level of detail, on the actions of an individual.

There has been some work on recognizing human actions using context information. Marszalek et al. [4] exploit scene-action context and demonstrate that recognizing the scene type (e.g. road scene) of a video helps the recognition of human actions (e.g. driving). Han et al. [5] uses object-action context, where the context of an action is implicitly defined by the objects (e.g. cars, pedestrians) detected in the scene. In this paper, we focus on another type of contextual information – the action-action context, i.e. the interactions between people. Modeling interactions between people and their role in action recognition has been explored by many researchers. For example, sophisticated models such as dynamic Bayesian networks [6] and AND-OR graphs [7] have been employed. Gupta et al. [7]'s representation based on AND-OR graphs allows for a flexible grammar of action relationships. The sophistication of these models leads to more challenging learning problems. Other representations are holistic in nature. Zhong et al. [8] examine motion and shape features of entire video frames to detect unusual activities. Mehran et al. [9] build a "bag-of-forces" model of the movements of people in a video frame to detect abnormal crowd behavior.

Choi et al. [10] consider spatial distributions of pedestrians and velocities from a tracker. Our representation for interactions has similarities to this work, but classifies actions rather than poses.

We apply this representation to a nursing home video dataset. As with many surveillance applications we are interested in finding rare actions – for example sifting through hours of footage to find the few instances of an action of interest. Hence, a standard action classification approach (e.g. SVM) is not appropriate. Instead, we employ rank-SVM [11] to rank video clips according to their degrees of relevance to a particular action query. We demonstrate that this is effective on the nursing home dataset.

The rest of the paper is organized as follows. Section 2 describes our feature representation. Section 3 introduces our rank-SVM based action retrieval method. We experimentally demonstrate the superiority of our method compared with other baseline approaches in section 4 and conclude in section 5.

2 Contextual Representation of Actions

Our approach enables analyzing human actions by looking at contextual information, which is extracted from the behaviour of all the people in a video frame. The input of our system is the raw video together with a set of hypothesized person locations in the video who might be performing the action of interest. Our goal is to retrieve or rank those people accordingly to their degree of relevance to the action of interest. How to localize people in the video frames is task-specific, and it involves either human detection [12] or background subtraction. We will describe the details in the experiments section. From now on, we assume the locations of people are given. We extract a feature vector from each person (section 2.1) that describes his shape or motion appearance. Then a contextual feature representation for each person is obtained by considering this person together with nearby people (section 2.2).

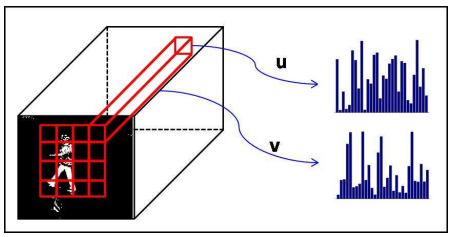


Fig. 2. Illustration of the local spatio-temporal (LST) feature representation for describing a person. **u** is a vector of percentage of static foreground pixels, **v** is a vector of percentage of moving foreground pixels

2.1 Person descriptor

For the nursing home dataset, standard features such as optical flow or HOG [13] are typically not reliable due to low video quality. Instead, we use a feature representation similar to the one introduced in [3], which has been shown to be reliable for low resolution videos. The feature descriptor is computed as follows. We first divide the bounding box of a detected person into N blocks. Foreground pixels are detected using standard background subtraction. Each foreground pixel is classified as either static or moving by frame differencing. Each block is represented as a vector composed of two components: $\mathbf{u} = [u_1, \ldots, u_t, \ldots, u_\tau]$ and $\mathbf{v} = [v_1, \ldots, v_t, \ldots, v_\tau]$, where u_t and v_t are the percentage of static and moving foreground pixels at time t respectively. τ is the temporal extent used to represent each moving person. As in [3], we refer to it as local spatio-temporal (LST) descriptor in this paper. Fig. 2 illustrates the LST descriptor.

2.2 Action context descriptor

We develop a novel feature representation called the *action context (AC) descriptor*. Our AC descriptor is centered on a person (the focal person), and describes the action of the focal person and the behavior of other people nearby. For each focal person, we set a spatio-temporal context region around him (see Fig. 3(a)), only those people inside the context region (nearby people) are considered. The AC descriptor is computed by concatenating two feature descriptors: one is the action descriptor that captures the focal person's action, and the other one is the context descriptor that captures the behaviour of other people nearby, as illustrated in Fig. 3(b,c).

Here we employ a bag-of-words style representation for the action descriptor of each person, which is built from a two-stage approach as follows. First, we train a multi-class SVM classifier based on the person descriptors (e.g. HOG [13] or LST introduced in Sec. 2.1) and their associated action labels. We then represent each person as a K-dimensional vector (i.e. the action descriptor), where K is the number of action classes. The action descriptor of the i-th person is: $F_i = [S_{1i}, S_{2i}, \ldots, S_{Ki}]$, where S_{ki} is the score of classifying the i-th person to the k-th action class returned by the SVM classifier.

Given the i-th person as the focal person, its context descriptor C_i is computed from the action descriptors of people in the context region. Suppose that the context region is further divided into M regions (we call "sub-context regions") in space and time, as illustrated in Fig. 3(b), then the context descriptor is represented as a $M \times K$ dimensional vector computed as follows:

$$C_i = \left[\max_{j \in \mathcal{N}_1(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_1(i)} S_{Kj}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{1j}, \ldots, \max_{j \in \mathcal{N}_M(i)} S_{Kj} \right] \quad (1)$$

Where $\mathcal{N}_m(i)$ indicates the indices of people in the m-th "sub-context region" of the i-th person.

The AC descriptor for the i-th person is a concatenation of its action descriptor F_i and its context descriptor C_i: $AC_i = [F_i, C_i]$. As there might be numerous

people present in a video sequence, we construct AC descriptors centered around each person. In the end, we will gather a collection of AC descriptors, one per person.

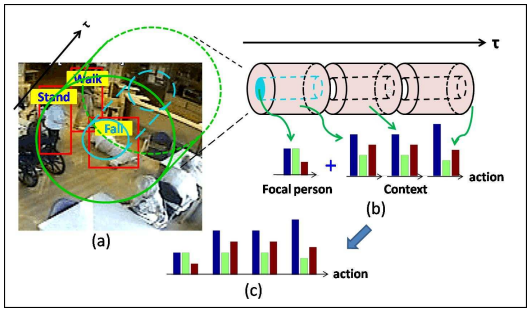


Fig. 3. Illustration of construction of our action context descriptor. (a) Spatio-temporal context region around focal person, as indicated by the green cylinder. In this example, we regard the fallen person as focal person, and the people standing and walking as context. (b) Spatio-temporal context region around focal person is divided in space and time. The blue region represents the location of the focal person, while the pink regions represent locations of the nearby people. The first 3-bin histogram captures the action of the focal person, which we call the action descriptor. The latter three 3-bin histograms are the context descriptor, and capture the behaviour of other people nearby. (c) The action context descriptor is formed by concatenating the action descriptor and the context descriptor.

Fig. 4 shows examples of the action context descriptors on the nursing home dataset. On this dataset, we label each person to be of the following six action classes: "walking", "sitting", "standing", "falling", "helping fallen residents to stand up", and "other". We use the last action class "other" to label person not belonging to any of the previous five categories or noise produced by background subtraction. Fig. 4(a) and Fig. 4(b) are two frames that contain "falling". The persons in the red bounding boxes are trying to help the fallen residents. Fig 4 is a frame that does not contain the falling action. The person in the red bounding box is simply walking across the room. For our application, we would like to retrieve examples in Fig. 4 (a,b), but not Fig. 4 (c). However, this is difficult (even for human observers) if we only look at the person in the bounding box, since all three people are walking. But if we look at the context of them, we can easily tell the difference: people in Fig. 4 (a,b) are walking to help the fallen residents, while the person in Fig. 4 (c) is simply walking. This can be demonstrated by the action context descriptors shown in Fig. 4 (d)-(f). Here use a 24-dimensional action context descriptor and visualize it as a 4×6 matrix so it is easier to compare them visually. We can see that Fig. 4 (d) and Fig. 4 (e) are similar. Both of them are very different from Fig. 4 (f). This demonstrates that the action

context descriptor can help us to differentiate people helping fallen residents from other actions, such as walking.

The key characteristics of our action context descriptor are in two aspects: 1) instead of simply using features of the neighboring people as context, the action context descriptor employs a bag-of-words style representation which captures the distribution of actions of people nearby. 2) In addition to static context, our descriptor also captures dynamic information, i.e. the temporal evolution of actions extracted from both the focal person and the people nearby.

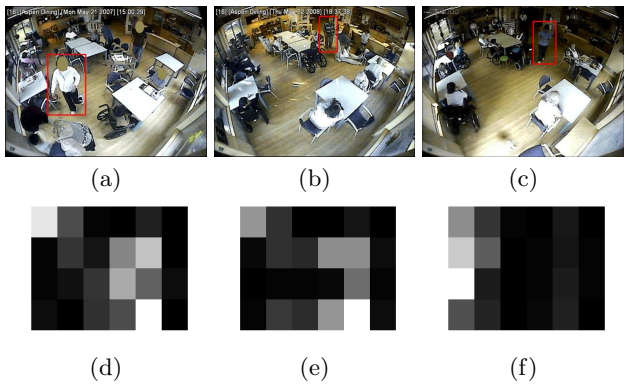


Fig. 4. Examples of action context descriptors. (a,b) Sample frames containing people falling and other people (shown in red bounding boxes) trying to help the fallen person. (c) A sample frame contain no falling action. The person in the red bounding box is simply walking. (d-f) The action context descriptors for the three persons in bounding boxes. Action context descriptors contain information about the actions of other people nearby.

3 Action Retrieval as Ranking

Most work in human action understanding (e.g. [14–17]) focuses on action classification. The goal is to classify a video as one of the pre-defined action categories defined on standard benchmark datasets, e.g. the KTH dataset [14], the Weizmann dataset [15]. In this work, we would like to argue that action classification is not necessarily the right problem formulation in understanding human actions in videos.

Thousands of hours of videos are being captured everyday by CCTV camera, web camera, surveillance camera, etc. However, most of the actions of interest only occur in a relatively small region along the spatial and temporal extent of the video. In this scenario, the task is typically to *retrieve* a small spatial/temporal segment of the video containing a particular action, rather than to *classify* the videos or the frames.

Our work on action retrieval is directly inspired by the application of fall analysis in nursing home surveillance videos. Our clinician partners are studying the causes of falls by elderly residents in order to develop strategies for prevention. This endeavor requires the analysis of a large number of video recordings of falls. Alternatives to vision-based analysis for extracting fall instances from a large amount of footage, such as wearable sensors and self-reporting, are inconvenient and unreliable.

Given a large collection of surveillance videos captured in nursing homes, the task is to retrieve video segments containing people falling. One straightforward solution to this problem is to classify each video segment to either "fall" or "non-fall". But there are two drawbacks with this approach. First, the two classes ("fall" versus "non-fall") are extremely imbalanced – falls are rare events. Second, there is a disconnect between classification and how the system will be deployed for use eventually. For clinicians, they expect the system to automatically *rank* all the videos according to how relevant they are to the falling action. Then they can manually examine the top-ranked videos and pick a certain number of relevant ones (i.e. those containing fall actions) according to their need. So it is more appropriate to solve the problem as a retrieval/ranking task rather than a classification task.

We formulate our retrieval task as follows. Our goal is to train a retrieval system that ideally ranks all the people from a video so that those containing falls are ranked higher. Our training data consist of a collection of people extracted from training videos. We manually label all the training examples as 3 ("very relevant"), 2 ("relevant") or 1 ("irrelevant"). Examples labeled "very relevant" correspond to people falling. Examples labeled "relevant" contain people in the context of a falling action, e.g. other people that help the fallen resident to get up. Other examples are labeled as "irrelevant". Given this labeled dataset, we use the rank-SVM [11] to learn a model that attempts to rank "very relevant" examples at the top of the list, and rank "irrelevant" examples at the bottom of the list.

Let $\mathcal{D} = \{(x_i, y_i) : 1 \leq i \leq N\}$ be a set of N training examples, where $y_i \in \{1, 2, 3\}$. Let $\mathcal{S}$ be the set of (i, j) pairs defined as:

$$\mathcal{S} = \{(i, j) : 1 \leq i \leq N, 1 \leq j \leq N, y_i > y_j\} \tag{2}$$

We use $\phi(x_i)$ to denote the feature vector extracted from x_i. In this paper, we use the action context (AC) descriptor described in Sec. 2.2 as the feature vector $\phi(x_i)$. For a new test instance x, the degree of relevance of x to the falling action is measured by a linear function $f_w(x) = w^\top \phi(x)$, where w is the model parameter to be learned. The rank-SVM learns the parameter w from the training dataset $\mathcal{D}$ by solving the following optimization problem:

$$\min_{w, \xi \geq 0} \ \frac{1}{2}||w||^2 + C \sum_{(i,j) \in \mathcal{S}} \xi_{i,j} \tag{3}$$

$$\text{s.t.} \quad w^\top \phi(x_i) \geq w^\top \phi(x_j) + 1 - \xi_{i,j} \quad \forall (i, j) \in \mathcal{S} \tag{4}$$

The intuition of Eq. 4 is as follows. For a pair $(i, j) \in \mathcal{S}$, by the definition of $\mathcal{S}$, x_i is more relevant to the query than x_j. So we would like w to score x_i higher than x_j by a margin of at least 1. This translates to:

$$w^\top \phi(x_i) \geq w^\top \phi(x_j) + 1 \tag{5}$$

Of course, there might not exist such w which satisfies Eq. 5 for all (i, j) pairs in $\mathcal{S}$, so we need a slack variable $\xi_{i,j}$ to handle the case of soft-margins. The parameter C in Eq. 4 is a trade-off parameter similar to that in regular SVMs.

The optimization problem in Eq. 4 is convex and can be solved by a cutting-plane algorithm [18].

4 Experiments

Most previous work in human action understanding uses standard benchmark datasets to test their algorithms, such as KTH [14] and Weizmann [19] datasets. In the real world, however, the appearance of human activities has tremendous variation due to background clutter, partial occlusion, scale and viewpoint change, etc. The videos in those datasets were recorded in a controlled setting with small camera motion and clean background. The Hollywood human action dataset [17] is more challenging. However, only three action classes: HandShake, HugPerson and Kiss have more than one actor, but these are not contextual – the 2 actors together perform the one action.(One person does not perform HugPerson by himself.) In this work, we choose to use two challenging datasets to evaluate our proposed method. The first dataset is a benchmark dataset introduced in [10] to study collective human activities. The second dataset consists of surveillance videos collected from a nursing home environment by our clinician collaborators.

(a) (b)

Fig. 5. Typical results of running a state-of-the-art pedestrian detector [12] on the two datasets used in the experiments. On the collective activity dataset (a), the detector performs very well. But on the more challenging nursing home dataset (b), the detector is not reliable since the videos are captured by a fish eye camera, so persons in the videos are not in upright positions. In addition, the video quality is very poor.

Our main focus is on action retrieval. The goal of a retrieval system (e.g. search engines like Google) is to rank the data according to their relevance to the

query and return the top-ranked instances. For retrieval tasks, the classification accuracy is not a meaningful performance measure, since users typically only care about the few top-ranked instances. In addition, since the majority of the instances are irrelevant to the query, a high classification accuracy can be trivially achieved by classifying all the instances to be irrelevant.

In this paper, we use the *Normalized Discounted Cumulative Gains* (NDCG) [20] to measure the performance of our action retrieval approach. We first give a brief introduction to this metric. For a list of instances sorted in descending order of the scores returned by a learned ranking model, the NDCG score at the k-th instance is computed as: $NDCG_k = \frac{1}{N_k} \sum_{i=1}^{k} \frac{2^{r(i)}-1}{\log(1+i)}$, where k is called a truncation level, N_k is the normalization constant to make sure the optimal ranking gets an NDCG score of 1, and $r(i)$ is the rating of the i-th instance. We set the rating of a very relevant instance as 3, a relevant instance as 2 and an irrelevant instance as 1. NDCG evaluates a retrieval system by only considering the ranking quality of the top-k instances returned by the system. The NDCG gain is discounted by a ranked position based factor $log(1 + i)$. Intuitively, the truncation level k corresponds to the number of instances that the users will look through before giving up. In our experiments, we report NDCG values at three different truncation levels, corresponding to top 5%, top 20%, and 100% percent of the total number of instances being ranked, respectively.

We compare our method of using rank-SVM and action context descriptor with three different baseline methods. The first baseline uses exactly the same action context descriptor, but uses a regular binary SVM as the learning method. This will demonstrate the advantage of using rank-SVM for retrieval tasks. We also compare with SVM/rank-SVM trained based on feature descriptors (e.g. HOG) without context, in order to demonstrate the advantage of our action context descriptor.

As indicated by Fig. 3, our action context descriptor is controlled by several parameters. Here we set them empirically to fixed values: we assume that in space, the context region centered around the focal person is divided into two regions, the radius of each region is proportional to the height of the focal person h, which are set to $0.5h$ and $2h$ respectively. In time, the context region is equally divided into three regions, each region has a temporal extent of two.

4.1 Collective Activity Dataset

This dataset contains 44 video clips acquired using low resolution hand held cameras. In the original dataset, all the persons in every tenth frame of the videos are assigned one of the following five action categories: crossing, waiting, queuing, walking and talking. We apply our method to retrieve each of the five actions. In order to demonstrate the effectiveness of our descriptor, we also perform standard classification on this dataset, so we can directly compare with the classification results reported in [10]. Similarly to [10], we apply the pedestrian detector in [12] to find all the people and report action retrieval/classification results on a per-person basis. The pedestrian detector performs very well on this dataset with only a few false positive detections (see Fig. 5(a)).

Action Retrieval: The query given to our action retrieval system is one of the action categories. The goal is to retrieve those people that are relevant to the given query. In order to evaluate the retrieval performance, we need the ground-truth label of each person indicating how relevant it is to a given query. We obtain the ground-truth labels for retrieval tasks from the action category labels (on a per-person level) provided by the original dataset as follows. For a given query (say "walking"), all the people labeled as "walking" in the original dataset are considered as "very relevant". A person is considered as "relevant" if it is not labeled as "walking", but there exists another person labeled as "walking" in the same video. Other people are considered as "irrelevant". The ground-truth labeling for other queries can be similarly obtained.

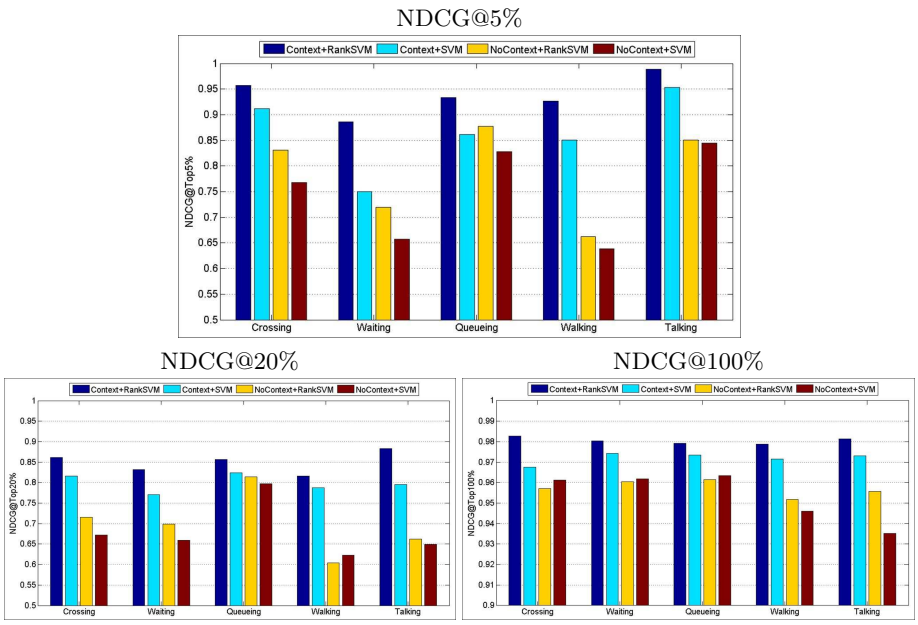


Fig. 6. Results of action retrieval on the collective activity dataset for each of the five action cateogires. We show NDCG values evaluated at three different truncation levels, corresponding to top 5%, 20% and 100% of all the examples, respectively.

We then split the dataset into six partitions and make sure that each action category exists in every partition. We use a six-fold cross validation scheme. At each run, we use one partition as the test set, and the other five partitions as the training set. We train a rank-SVM for each of the five queries (crossing, waiting, queuing, walking and talking) and measure the NDCG scores. The results are shown in Fig. 6. We can see that our proposed approach (context + rank-SVM) yields higher NDCG performance than the other baseline methods for all the queries. We can also see that the gap between our method and the baselines

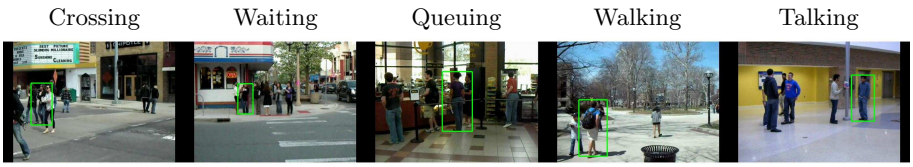


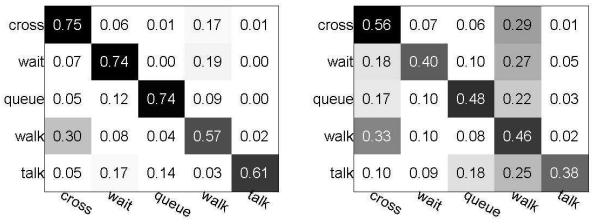
Fig. 7. Visualization of the retrieval results on the collective activity dataset. Here we show the most relevant instance (instance with the highest ranking score) for each of the five action classes (cross, wait, queue, walk and talk). The color of the bounding box indicates the ground-truth label for a particular retrieval task. Green represents "very relevant" in ground truth, blue represents "relevant", red represents "irrelevant". We can see that the topmost-ranked instances for the five classes are all instances labeled as "very relevant".

in terms of NDCG scores is larger when considering the top examples (e.g. 5%, 20%). This is very desirable since users typically cannot sift through all the retrieved results. Instead they will probably focus on the top few percent of the retrieved results. Fig. 7 visualizes the top-ranked instances during one of the 6 runs.

Action Classification: We perform standard action classification in order to evaluate the proposed HC descriptor separately from the learning scheme. Note that though the focus of this paper is action retrieval, our HC descriptor can also be used for action classification. In this way, we can demonstrate the strength of the proposed HC descriptor by comparing it with what proposed in [10] on a benchmark dataset. In order to make a fair comparison with [10], we use the same leave-one-out scheme described in [10]. When classifying people's actions in one video, we use people from all the other videos as the training set. As a baseline comparison, we also report the result of a multi-class SVM classifier on HOG features extracted from the person. The confusion matrix of our method and the baseline are shown in Fig. 8. We also compare our method with the spatio-temporal descriptor in [10] trained with SVMs. Since the classifier is identical, the comparison is fair. We summarize the comparison in Table 1. Please note that Table 1 also lists the best result reported in [10]. However, since this accuracy number is achieved by using other information such as velocity, it is not directly comparable to other numbers listed in the table. We can see from Table 1 that our HC descriptor outperforms other feature representations without context in the action classification task.

4.2 Nursing Home Data

Our second dataset consists of videos recorded in a dining room of a nursing home by a low resolution fish eye camera. Typical activities happening in nursing homes include people walking, sitting, standing, falling and people helping the fallen person to stand up, etc. Our dataset contains ten 3-minutes video clips without falls and another ten short clips with falls. We demonstrate the retrieval



(a) (b)

Fig. 8. Confusion matrices for action classification on the collective activity dataset using SVMs with two different feature representations: (a) AC descriptor (b) HOG.

Method	Accuracy
action context descriptor + SVM	**68.2**
HOG + SVM	45.6
spatio-temporal descriptor + SVM in [10]	57.4
best result in [10]	65.9

Table 1. Comparison of classification accuracies of different methods on the collective activity dataset. Now the best result in [10] is achieved by using other information such as velocity of people. So it is not directly comparable to other numbers in the table.

of people falling on this dataset, since this is the most interesting and relevant action for clinicians.

Since pedestrian detectors are not reliable on this data, we instead extract moving regions from the videos as our detected people. First, we perform background subtraction using the OpenCV implementation of the standard Gaussian Mixture Model (GMM) to obtain the foreground regions. Then, we extract all the 8-connected regions of the foreground from each frame, which are considered as moving regions. Moving regions with size less than a threshold Th are deemed unreliable and therefore ignored. We manually label all the people performing the action falling as "very relevant". It is very common that when a falling event happens, other people will try to approach the fallen person and help him/her to get up. The activities of those people provide a useful contextual cue to detect falling actions. We label those people as "relevant". All the other people are labeled as "irrelevant".

We use a 10-fold cross validation scheme to evaluate the performance. We split the dataset into 10 partitions. Each partition contains one clip without falls and one clip with falls. During each run, we use one partition as the test set and the other nine partitions as the training set. The results are shown in Fig. 9. Similar to the results on the collective activity dataset, our method (context+rank-SVM) outperforms other baseline methods. Fig. 10 visualizes the top-ranked and bottom-ranked instances during one of the 10 runs. The green bounding box means the ground-truth label of the person is "very relevant", the blue bounding box means "relevant", the red bounding box means "irrelevant".

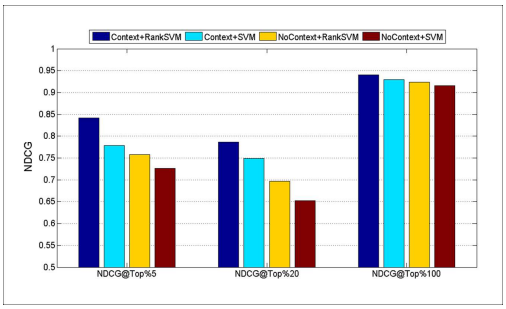


Fig. 9. NDCG values evaluated at three different truncation levels, corresponding to top 5%, 20% and 100% of all the examples, respectively.

We can see that the top-ranked instances are all instances labeled either as "very relevant" or "relevant". This demonstrates that our approach provides a useful tool for clinicians to quickly retrieve falling actions in those surveillance videos.

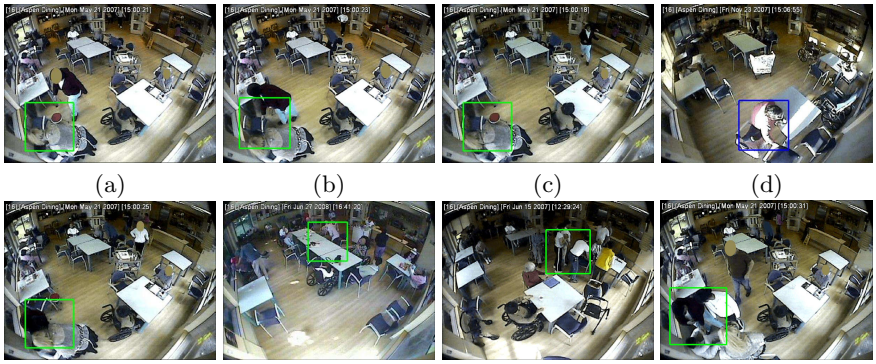


Fig. 10. Visualization of the retrieval results on the nursing home dataset. (a)-(h): the top 8 instances from ranking. Green bounding box represents "very relevant" in ground truth, blue represents "relevant", red represents "irrelevant". Note that in some examples, the fallen residents are occluded by the people coming to help (e.g. (a)-(c)), we could still retrieve them by the contextual information (e.g. people helping the fallen residents to stand up).

5 Conclusion

We have developed methods for action retrieval from surveillance videos using contextual feature representations. Our proposed AC descriptor encodes information about action of an individual person in a video, as well as behaviour of other people nearby. Our experimental results demonstrate the advantage of

using contextual information when dealing with complex activities. Another contribution of this work is to introduce the action retrieval formulation, which is different from previous action classification work. We address the video retrieval task using rank-SVM, a state-of-the-art learning technique specifically designed for retrieval tasks, but has not been deployed extensively in the computer vision community. In our results, rank-SVMs outperforms regular SVMs.

References

1. Wang, Y., Mori, G.: Human action recognition by semi-latent topic models. IEEE Trans. PAMI **31** (2009) 1762–1774
2. Wang, X., Ma, X., Grimson, E.: Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models. IEEE Trans. PAMI **31** (2009) 539–555
3. Loy, C.C., Xiang, T., Gong, S.: Modelling activity global temporal dependencies using time delayed probabilistic graphical model. In: ICCV. (2009)
4. Marszalek, M., Laptev, I., Schmid, C.: Actions in context. In: CVPR. (2009)
5. Han, D., Bo, L., Sminchisescu, C.: Selection and context for action recognition. In: IEEE International Conference on Computer Vision. (2009)
6. Xiang, T., Gong, S.: Beyond tracking: Modelling activity and understanding behaviour. Int. Journal of Computer Vision **67** (2006) 21–51
7. Gupta, A., Srinivasan, P., Shi, J., Davis, L.S.: Understanding videos, constructing plots - learning a visually grounded storyline model from annotated videos. In: CVPR. (2009)
8. Zhong, H., Shi, J., Visontai, M.: Detecting unusual activity in video. In: Proc. IEEE Comput. Soc. Conf. Comput. Vision and Pattern Recogn. (2004)
9. Mehran, R., Oyama, A., Shah, M.: Abnormal crowd behavior detection using social force model. In: CVPR. (2009)
10. Choi, W., Shahid, K., Savarese, S.: "what are they doing? : Collective activity classification using spatio-temporal relationship among people". In: VS. (2009)
11. Joachims, T.: Optimizing search engines using clickthrough data. In: ACM SIGKDD. (2002)
12. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: CVPR. (2008)
13. Dalal, N., Triggs, B.: Histogram of oriented gradients for human detection. In: CVPR. (2005)
14. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: A local svm approach. In: 17th International Conference on Pattern Recognition. (2004)
15. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: Proc. 10th Int. Conf. Computer Vision. (2005)
16. Niebles, J.C., Wang, H., Fei-Fei, L.: Unsupervised learning of human action categories using spatial-temporal words. In: BMVC. (2006)
17. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: CVPR. (2008)
18. Joachims, T.: A support vector method for multivariate performance measures. In: International Conference on Machine Learning. (2005)
19. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: ICCV. (2005)
20. Chapelle, O., Le, Q., Smola, A.: Large margin optimization of ranking measures. In: NIPS Workshop on Learning to Rank. (2007)