Unsupervised classification and visual representation of situations in surveillance videos using Slow Feature Analysis for situation retrieval applications

Frank Pagel
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation IOSB
Dept. of Video Exploitation Systems VID, Fraunhoferstr. 1, 76131 Karlsruhe, Germany

ABSTRACT
Today, video surveillance systems produce thousands of terabytes of data. This source of information can be very valuable, as it contains spatio-temporal information about abnormal, similar or periodic activities. However, a search for certain situations or activities in unstructured large-scale video footage can be exhausting or even pointless. Searching surveillance video footage is extremely difficult due to the apparent similarity of situations, especially for human observers. In order to keep this amount manageable and hence usable, this paper aims at clustering situations regarding their visual content as well as motion patterns. Besides standard image content descriptors like HOG, we present and investigate novel descriptors, called Franklets, which explicitly encode motion patterns for certain image regions. Slow feature analysis (SFA) will be performed for dimension reduction based on the temporal variance of the features. By reducing the dimension with SFA, a higher feature discrimination can be reached compared to standard PCA dimension reduction. The effects of dimension reduction via SFA will be investigated in this paper. Cluster results on real data from the Hamburg Harbour Anniversary 2014 will be presented with both, HOG feature descriptors and Franklets. Furthermore, we could show that by using SFA an improvement to standard PCA techniques could be achieved. Finally, an application to visual clustering with self-organizing maps will be introduced.

Keywords: unsupervised learning, crowd behaviour analysis, Slow Feature Analysis, Franklets, situation classification, motion patterns, PCA, SOM

1. INTRODUCTION AND RELATED WORK
Surveillance tasks are increasingly augmented with vision systems and smart algorithms in order to detect events and also find events in large data bases - preferably in real-time. In this work, we focus on the definition of useful features for spatio-temporal situation modelling and retrieval. Hereby, several challenges arise, such as strong visual similarities of crowded scenes, motion estimation and dependencies of object tracker models, loss of tracks, environmental conditions like weather or lights, crowded scenes, noise, etc. The objective is to support situation retrieval for surveillance footage by clustering and representing situations appropriately.

In this paper, a situation is characterized by its spatial and temporal information. A spatial dependency of image regions is simply reached by dividing the image into cells and by processing each cell independently\(^\ast\). This cell-wise implementation already enables the definition of simple AND queries, like *Show all situations, where situation S1 happened in cell C1 and situation S2 in cell C2.* Texture as well as motion information is important for a human analyst as it helps navigating intuitively through a video. Therefore, in this paper we used Histograms of Oriented Gradients (HOG)\(^1\) as textural image descriptors. Motion features should be able to answer questions like *Where did objects come from, that passed image cell C?* and *Where did object go to after passing cell C?* Therefore, novel descriptors called Franklets are introduced in this paper, which explicitly encode motion patterns for certain image regions and is based only on dense optical flow fields.\(^2\) Additionally, Franklet descriptors also encode the motion information visually (see Figure 1).

In many contexts, it would be desirable to learn situations in an unsupervised manner. For that reason we chose Self-Organizing Maps (SOMs)\(^4\) to cluster events of each cell independently. In SOMs, similar features always regionally closely organised and hence is a good way to present events to human analysts in a visual way. In order to keep the

\(^\ast\) Future work will aim at connecting these regions and finding relations between the cells.
data manageable, we seek for dimension reduction possibilities. Besides the classical Principal Component Analysis (PCA) approach, we considered Slow Feature Analysis (SFA), which has its advantage in classifying temporal features.

Event detection and unsupervised clustering of activities in video sequences is a very wide, not to say classical research field in computer vision. Hence, many different approaches were proposed during the last years in different contexts. Zhou et al. for example classified events in facial expressions, while Niebles at al. or Zhang et al. detected human activities in video data. Schuster et al. tried to detect abnormal behaviour and anomalies in surveillance video footage based on HOG-like features and an unsupervised clustering algorithm which is an online version of the classical K-Means algorithm, called Meaningful K-Means. While flexible and effective due to its simple rule-based implementation for each cell of an image grid, this approach does not consider motion explicitly.

Many approaches used motion as the basic criteria for event classification in surveillance videos. Therefore, mostly trajectories of persons or objects were used. Stauffer and Grimson used a background model in order to extract objects and track them in the image. Others, like learned statistical models of object trajectories. These trajectory-based approaches suffer from error-prone trackers as usually it is very challenging to find a tracker that can be applied to different scenarios and camera configurations. So, in order to yield good results, a specialized tracker would be needed, which would make the system unflexible for a usage at large variety of surveillance data from several video cameras. Also, those tracker-based approaches have difficulties when it comes to dense crowds of people as tracks often get lost or targets may be switched due to object similarities, especially in large crowds.

SFA was used in recent publications in order to achieve a dimension reduction which minimizes temporal variations and to take the temporal dependencies of video events into account. SFA features are suitable for learning stable representations of activities, situations or human behavior. E.g. Zhang et al. recently investigated SFA for human action recognition.
Figure 2. Dense Optical Flow field from a sequence of the Hamburg Harbour Anniversary 2014. The motion direction is color-coded. The water was masked in order to suppress noise caused by water reflections. The rectangle marks the cell that was used for the evaluations in Section 3.

Theriault et al.\textsuperscript{14} presented a SFA-based framework for classification of TV video sequences and Nater et al.\textsuperscript{15} proposed a combination of PCA and SFA for learning abnormal human activities in an industrial working environment.

In this paper we use texture and motion features in order to describe and classify activities in image cells. We use HOG features as well as Franklet features, that will be described in more detail in section 3. Furthermore, these features will be processed with PCA and SFA in order to yield more discriminative and lower-dimensional features. Then, these features will be clustered in an unsupervised manner, so that similar situations can be grouped into categories. Feature processing and clustering finally are the base for retrieval applications, which support analysts in detecting anomalies, similar situations and can hence initiate further retrieval activities (e.g. object retrieval\textsuperscript{16} or person retrieval\textsuperscript{17} tools).

The online algorithm for situation classification and representation can be summarized as follows: For all image cells do: 1) Calculation of video features (HOGs, Franklets) 2) Calculation of PCA (noise reduction) 3) Calculation of SFA features 4) Map into SOM and visualize. An evaluation on real video data from the Hamburg Harbor Anniversary 2014 (1.5 million visitors in 3 days) is used in this paper for investigating the suitability of the resulting features of classification and hence for online learning and retrieval tasks. These results will be presented in section 4.

\section*{2. SPATIO-TEMPORAL SITUATION MODEL AND FEATURES}

In order to be independent of camera models, perspectives and specialized trackers, we resigned an approach based on object tracking. Instead, we developed generic motion patterns based on dense optical flow fields (see Figure 2), we call Franklets. For that purpose, the image is divided into a grid of cells. A Franklet for cell $c_i$ is calculated by accumulating flow vectors that have their origin in $c_i$ over a history of $H$ frames. This results in tracks for each pixel in $c_i$ which in turn can be used to calculate occurrence histograms over the image. This 2D-distribution describes where and how long a pixel in $c_i$ also occurred in other image regions and hence encodes information about where objects came from before they reached $c_i$. It is one advantage of these histograms that they already visually describe motion patterns for each cell (for example for a visual representation for human analysts). Figure 1 shows Franklet motion histograms for some cells in a grid of size $6 \times 5$. A Franklet feature vector is finally generated by stacking the rows of the 2D-histogram.
Franklets can be calculated in a backward as well as a forward manner. The latter one gives hints about the image regions where objects moved to after they passed $c_i$. We call the backward calulation “P-Franklet” (Past) and the forward calculation “F-Franklet” (Future).

In addition to Franklet features, HOG features are used to describe cell textures.

Features extracted directly from the image data are here called “Level 1” ($L_1$) features. Both, Franklet and HOG feature vectors are high-dimensional, usually $\dim \gg 100$. As a next step, we want to cluster and visualize textural and temporal features in lower dimensional feature spaces. Therefore, Slow Feature Analysis (SFA)\textsuperscript{6} will be applied as used and described in more detail by Theriault et al.\textsuperscript{14} or Nater et al.\textsuperscript{15} The idea of SFA is that temporally adjacent features (e.g. in a sequence between two or more consecutive frames) are likely to be correlated and to belong to the same class. So, in dependence of the input features’ temporal covariance matrices, SFA determines the temporally least varying and hence most stable features. By choosing only the $m$ “slowest” dimensions, lower dimensional features with higher discriminative properties can be achieved. Figure 3 illustrates the franklet features vectors before and after applying a SFA transformation.

We compared SFA dimension reduction to a standard Principal Component Analysis (PCA) approach.\textsuperscript{5} PCA determines and transforms the coordinate system according to the highest variance of the features’ distribution in the feature space. We also implemented a hybrid approach which is a combination of SFA and PCA, where we trained a SFA model on the 20 first dimensions after a PCA, as PCA is can be a useful intermediate step for noise reduction.\textsuperscript{15} Both approaches, SFA and PCA, can be learned in an unsupervised manner, but require an appropriate training data set.

Due to the drastic dimension reduction by SFA, situations can be visualized according to their similarity. This can be done by clustering them fast and efficient in an unsupervised manner with Self-Organizing Maps (SOMs).\textsuperscript{14} SOMs can map higher dimensional features according to their similarity onto a two-dimensional map. This provides an intuitive representation for human analysts. For example, each node in the SOM can be linked with a representative thumbnail of the cell that is visualizing the image cell as some kind of “scene snippet”. So in practice, a human operator would be able to select one (ore several) cells, select one out of several feature modes (e.g. HOG, Franklet Forward/Backward or others) and then select a situation class in the SOM. This approach can be ideally used for supporting semi-automatic search tasks and navigation in large-scale video surveillance footage.

3. EXPERIMENTS AND RESULTS

For developing, training and evaluation of the proposed algorithms, data sets from the Hamburg Harbor Anniversary 2014 were used, where we had the possibility to capture video data for two complete days from a high-rise building. Our Franklets had a dimension of $20 \times 16 = 320$, the HOG features $\dim = 441$. All features were calculated and clustered for the cell marked in Figure 4. In a first step the optical flow field was calculated, and based on that, the $L_1$-features: P-Franklets ($ID_0$), F-Franklets ($ID_1$) and HOGs ($ID_2$). We used a history for the calculation of the Franklets $H = 100$. For the training of the SFA and PCA parameters, two sequences with a length of two hours each were used. Then, another three test sequences were used (two hours each) where we calculated the $L_1$-features and projected them into the SFA- and PCA-spaces respectively. The resulting features had a dimension of 3 ($SFA_3$ and $PCA_3$). For the hybrid approach, we first used the first 20 PCA dimensions with the highest variance ($PCA_{20}$) and projected this 20-dimensional feature vector into the SFA space, where we again only took the three slowest features ($SFA_3(PCA_{20})$). Figure 4 shows the effects of dimension reduction via SFA and PCA.

The reason for choosing $\dim = 3$ is that these features can be visualized in a very simple and effective way. In order to cluster and classify features in an online and unsupervised manner we implemented a SOM with 100 $\times$ 100 neurons. This means that also in real-time applications new situations can be added, classified and re-clustered. SOMs cluster features in a way that similar features are mapped close to each other. With $\dim = 3$ the weight of each feature can be color-coded, e.g. visualized in RGB-space. This visualization is very intuitive as it immediately allows for distinguishing different and similar features and hence situations (see Figure 5). This map is especially useful for human analysts in order to find similar features and situations fast and efficiently. However, as can also be seen in Figure 5, it turns out that HOG features - at least for the given data set - are not very useful when it comes to discriminating different textures in crowded events. Here, other features like color histograms or crowd density estimators might be more suited.

In order to investigate the eligibility of the features for clustering, a K-Means cluster algorithm was executed on the test data sets. The clustering was performed for the “original” $L_1$ features as well as for the projected and hence dimension reduced features ($SFA_3$, $PCA_3$, $SFA_3(PCA_{20})$), each for P-Franklets, F-Franklets and HOGs, respectively.
Figure 3. P-Franklets from another Mall-sequence. At the bottom, the input signal is displayed in a bin representation (left) and over time (middle) as well as the 3 best SFA features (right). Top left, there is a scheme from which illustrates the principle of SFA.

Figure 4. Screen shot from the Hamburg-sequence and the calculated features (the diagrams $SFA_3$, $PCA_3$ and $SFA_3(PCA_{20})$ are plotted over time).

Figure 5. Unsupervised clustering with SOMs ($100 \times 100$). Each 3-dimensional neuron weight is RGB-color-coded. The crosses indicate the cluster centers as resulted from the K-Means algorithm ($K = 20$). Each pixel in the map represents a feature class.
As a measure for the quality of the resulting clusters we used the compactness $\sum_i^M \| f_i - \text{ctr}(l_i) \|^2$, where $M$ is the number of all features, $f_i$ is the $i$th feature, $l_i$ is the cluster label of $f_i$, and $\text{ctr}(l_i)$ the respective cluster center. The compactness describes how close the features lie to their respective cluster center and hence describes the quality of the distribution of the feature points in the feature space.

Figure 7 shows the results of the K-Means evaluation for different values of $K$. It is clearly visible that the hybrid approach $SFAG_3(PCA_{20})$ yields the most compact cluster results. This is because the PCA works like some kind of low-pass filter. But it can be seen that even $SFAG_3$ yields a higher compactness than $PCA_3$.

There was no ground truth for the Hamburg sequences available that could be used for a quantitative evaluation and comparison of the cluster results. Therefore, Figure 6 exemplarily shows F-Franklet icons from two classes resulted from the K-Means analysis ($K = 20$) on the $SFAG_3(PCA_{20})$ data. One can intuitively see by comparing the motion histograms of the Franklets that similar motion patterns are mostly clustered in a meaningfully way after SFA was applied. However, problems may arise when the training data set differs too much from the test data. Because then, SFA and PCA parameters cannot be applied properly to the incoming data or do not result in intuitive cluster results. Also, too complex scenes can cause artifacts in the motion patterns, for example occlusions, that disrupt tracks and flow fields or extreme lighting conditions. In such situations, additional features need to be taken into consideration in order to describe the spatio-temporal dynamics of the scene adequately.
Figure 7. Cluster result: Compactness with K-Means for different values of $K$. Top row: Compactness of P-Franklet ($ID_0$) and F-Franklet ($ID_1$) features. Each diagram shows the compactness of the clustered $L1$-feature, the projected $PCA_3$ feature, as well as the $SFA_3$ and $SFA_3(PCA_{20})$ features. Middle row: Comparison of the investigated dimension reduction approaches for P- and F-Franklet features. For both feature types, the hybrid approach $SFA_3(PCA_{20})$ performs best. Bottom row: Same analysis as above for HOG features ($ID_2$).
4. CONCLUSION

In this paper, a grid-based framework for learning and representing situations for the purpose of situation retrieval was presented. It could be shown that by using SFA a significant improvement of the clustering results was achieved compared to our baseline evaluation data set based on standard PCA techniques. Novel features called Franklets have been introduced, that can be used for unsupervised clustering as well as visual representation of motion patterns in videos. We demonstrated the practical use of Franklet descriptors for describing temporal events in order to support browsing applications in surveillance video footage. In that context, the hybrid approach of PCA and SFA could yield the most compact cluster results on the real data set from the Hamburg Harbour Anniversary 2014. The feature clustering algorithms were integrated into an application that visually clusters SFA-dimension-reduced features based on self-organizing maps (SOMs). It also turned out, that HOG features have only little practical use when it comes to describing textures of highly crowded scenes.

Of course, in order to describe situations in their whole complexity, more features need to be added to the framework, such as color histograms, crowd density or others. The selection of features depends on the specific use case. So, further work in this area will focus on the usage of the presented features for retrieval applications in large video archives of surveillance footage. Also, more features that are suited for situation modelling will be integrated. This also involves the implementation of a graphical user interface for an intuitive navigation in the footage and the retrieval results as well as the interface to other existing retrieval software modules. Furthermore, an additional software module will be developed in order to be able to perform situation retrieval for several image cells at once.

REFERENCES
