Honda Research Institute EU
Honda Research Institute Europe GmbH
https://www.honda-ri.de/

Non-negative Sparse Coding for Motion Extraction

Thomas Guthier, Volker Willert, Karel Kreuter, Andrea Schnall, Julian Eggert

2013

Preprint:

This is an accepted article published in Int. Joint Conf. on Neural Networks (IJCNN). The final authenticated version is available online at: https://doi.org/[DOI not available]

Non-negative Sparse Coding for Motion Extraction

T. Guthier, V. Willert, A. Schnall, K. Kreuter and J. Eggert

Abstract— **Visual motion is a rich source of information that is directly coupled to the underlying shape of a moving object. One way to describe motion is to use optical flow fields. Due to the aperture problem, dense optical flow estimation is an ill-constraint problem, while sparse optical flow estimation looses the shape information of moving objects. Current estimation algorithms based on regularization or segmentation fail at surface deformations or when the relevant motion is less dominant then its sourrounding movements. Both is e.g. true for face movements, where small movement patterns, so called action units, need to be preserved for further image analysis. We present a novel approach to capture the characteristics of local motion patterns that is based on the brightness constancy equation of optical flow estimation in combination with feature extraction using translation invariant non-negative sparse coding. Our approach simultaneously learns basic motion patterns and estimates the flow field without requiring pretrained motion patterns from ground truth optical flow data. We show on a face expression dataset how this method can preserve weak movements even in the presence of large head movements.**

I. Introduction

Visual motion is a rich source of information for animals as well as for artificial vision systems in a variety of fields, such as human-human or human-machine interaction, gait recognition, human action recognition and face expression recognition (FER).

Due to a variety of different facial movements and deformations FER is a particularly hard problem for motion estimation. The task in FER is to recognize the emotions of a person based on its visual expressions and facial movements. In [13] the psychologist Paul Ekman claims that 6 basic emotions (*anger*, *fear*, *happiness*, *sadness*, *disgust* and *surprise*) can be detected based on so called *action units* (AU), possibly small but specific movements of facial regions, such as the outer parts of the lips or the eyebrows, etc.

Current methods that try to estimate basic emotions mostly rely on appearance features and spatial relations rather than on motion-based features [15]. The reason for this is that there are no reliable methods to capture the relevant motion patterns. Sparse feature-point based approaches depend either on manual feature point labeling or reliable feature point detectors. Besides, feature-point approaches cannot capture local form information related to the movement of the AU. On the other hand, dense approaches, like dense *optical flow estimation* (OFE), discribe the movement of every pixel, thus they implicitly represent shape and motion simultaneously. Since OFE is an ill-constraint problem [5], current OFE algorithms either use spatial regularization or segmentation methods. Both approaches fail if the observed object is either deformable and thus not easily segmentable or if the relevant motion is small relative to its sourrounding motion. Both is true for face movements.

In this work, we present an unsupervised learning algorithm that captures small motion patterns, despite global motion and deformations, without any preprocessing or feature labeling. Our approach is based on the idea of the existence of a finite set of re-occuring basic motion patterns that can describe the relevant neighbouring relations for a large set of applications and natural movements. Examples for these patterns could e.g. be the facial AU or human body movements. The proposed algorithm is able to learn these motion patterns and detect them in the image in a simultaneous manner.

Unsupervised learning algorithms like *sparse coding* (SC) [12] or *non-negative matrix factorization* (NMF) [9] are often used to learn a set of basic patterns out of a set of given data. The basic patterns learned with these algorithms yield similar responses to natural image stimuli as those found in the early human visual cortex and the extracted patterns can be interpreted as oriented edge or gradient detectors. In a recent publication [4] SC is applied to video data to learn sparse motion patterns. Unlike images, motion fields contain negative values. In [4] a complex valued representation is applied and motion is interpreted via the phase shifts of complex valued patterns. In our own previous work [1] we preserve the non-negativity condition by using a direction selective representation of the motion field. This approach is neurophysiologically more plausible, due to the direction selectiveness of the patterns found in the motion processing pathway [17], as well as the non-negative nature of neuronal activity. A side benefit of a non-negative encoding is that we can apply the fast and parameter free multiplicative gradient descent update rules of the NMF [9] to extract the motion features.

In computer vision, learning methods are already used to overcome the ill-constrained dense OFE problem. Dense OFE was originally introduced by Horn and Schunk [5], who solved the motion estimation problem by minimizing the *brightness constancy equation* (BCE) with an additional regularization function, that enforces neighbouring pixels to have a similar motion vector. This regularization constraint leads to smoothing of edges as well as to error-prone behavior in unstructured image regions. Recent publications mainly focus on robust optimization methods and edge preserving, segmentation-based approaches [7]. In [2], [3] the additional constraint to the BCE is that the flow field around each pixel should be reconstructable by a local sparse code [2] or a

T. Guthier, V. Willert, A. Schnall and K. Kreuter are with the Control Theory and Robotics Department, Technische Universitaet Darmstadt, Germany (email: {tguthier, vwillert, aschnall, kreuter}@rtr.tu-darmstadt.de).

J. Eggert is with the Honda Research Institute Europe, Germany (email: Julian.Eggert@honda-ri.de).

set of principal components [3]. The sparse basic patterns as well as the principal components have to be pre-learned on an optical flow dataset.

Our approach is fundamentally different, in the sense that we use a translation invariant non-negative SC [10], [11] that allows us to reconstruct the entire flow field and not only the local sourrounding of each pixel. Rather than an additional condition, we directly force the flow field to consist of the linear superposition of the basic flow patterns by inserting the reconstruction directly into the BCE. We can thus learn the basic patterns and estimate the flow field simultaneously. We show on the challenging MMI face expression database [14] that our method preserves small motion patterns as those needed for the detection of AU. We further present a simple setup how to use the flow field to distinguish between the 6 basic emotions from [13].

II. Algorithm

The goal of our algorithm is to reconstruct the movement of every pixel between two consecutive frames via a superposition of translation invariant basic motion patterns. Thus, we start with the linearized BCE

$$E_{BCE,i} = \frac{1}{2}\|\mathbf{I}_{i,x} \circ \mathbf{V}_{i,x} + \mathbf{I}_{i,y} \circ \mathbf{V}_{i,y} + \mathbf{I}_{i,t}\|_2^2. \quad (1)$$

with $\circ$ being the element-wise multiplication of matrices; $\mathbf{I}_{i,x}$ and $\mathbf{I}_{i,y}$ are the spatial derivations in x- and y-direction for the ith image, $i \in \{1, ..., N\}$, with N = number of input images. $\mathbf{I}_{i,t}$ is the temporal derivative and $\mathbf{V}_{i,x}$ and $\mathbf{V}_{i,y}$ describe the unknown movement of all pixels of the ith image in horizontal and vertical direction.

The BCE (1) is known to be ill-constrained [5] due to the aperture problem, so additional conditions are required to solve the motion estimation problem. Our constraint is that the motion vectors can only consist of a linear superposition of shift-invariant spatially extended basic motion patterns. Each pattern $\mathbf{W}_j$, $j \in \{1, ..., J\}$, with J = number of basic patterns, describes a neighbourhood around pixel $\mathbf{m}$ of the image I_i with a common translational movement. The direction and velocity is given by its corresponding two dimensional activity $h_{ij,f}(\mathbf{m})$, with $f \in \{x, y\}$. We get

$$\begin{aligned}\mathbf{V}_{i,f} \approx \mathbf{R}_{i,f} &= \sum_{j,\mathbf{m}} \mathbf{R}_{ij\mathbf{m},f} = \sum_{j,\mathbf{m}} h_{ij,f}(\mathbf{m}) \left(\mathcal{T}^{(\mathbf{m})}\mathbf{W}_j\right) \\ &= \sum_j \mathrm{conv}_2(\mathbf{H}_{ij,f}, \mathbf{W}_j), \end{aligned} \quad (2)$$

with $\mathcal{T}^{(\mathbf{m})}$ being a transformation matrix that shifts the basic vector $\mathbf{W}_j$ by $\mathbf{m}$ pixels and conv_2 is the two-dimensional convolution.

In related work [2], [12], [9], [10] the reconstruction is e.g. learned by minimizing the Euclidean reconstruction error

$$E_{rec} = \frac{1}{2}\|\mathcal{V} - \mathcal{R}\|_F^2 \quad (3)$$

for a given input matrix $\mathcal{V}$, where the inputs consist of small local image patches. Since we allow translations throughout the entire image, we can reconstruct the entire flow field at once. This enables us to include the reconstruction $\mathbf{R}_{i,f}$ from (2) directly into the BCE (1) leading to the energy function

$$E_{BCE,i} = \frac{1}{2}\|\mathbf{I}_{i,x} \circ \mathbf{R}_{i,x} + \mathbf{I}_{i,y} \circ \mathbf{R}_{i,y} + \mathbf{I}_{i,t}\|_2^2. \quad (4)$$

A. Non-negative Representation and Sparsity

Our constructed optical flow field consists of a linear superposition of each possible shift of each basic pattern. That is, every pixel of the flow field $\mathbf{R}_i$ consists not necessarily of just one, but of multiple translational motion patterns. This degree of freedom allows the construction of almost arbitrary flow fields.

To get a significant motion description we have to imply further constraints on our flow field. The local movements of pixels on rigid objects can be approximated by a common translational motion. Hence, if we restrict our basic patterns locally, one basic pattern alone should be sufficient to describe its part of the flow field. However, a sufficiently large number of different basic shapes $\mathbf{W}_j$ that can describe the occuring object boundaries is required.

In conclusion our basic patterns should be *small*, *locally connected* and *parts-based*. In addition and contrary to the local nature of the desired patterns, the patterns should be as spatially expanded as possible to overcome the aperture problem. We enforce the above mentioned conditions by choosing a *non-negative* representation with *sparsity* constraints on the activities.

The non-negativity constraints on the activities and basic patterns enforce, that components can only be added to the reconstruction and not substracted. Thus, if a basis vector $\mathbf{W}_j$ is used for a reconstruction $\mathbf{R}_i$ (i.e. $h_{ij}(\mathbf{m}) > 0$), it has to describe its specific part of the input, because there is no possibility to remove parts of its partial reconstruction $\mathbf{R}_{ij\mathbf{m}}$. Without non-negativity an overlapping second pattern could be substracted, leading to less meaningfull and more arbitrary basic patterns. NMF is generally known to give parts-based reconstruction [9], [10], [11], [1]. The additional sparsity constraint on the activities penalizes the use of multiple activities and thus further enforces a parts-based reconstruction. Sparsity is also necessary to circumvent trivial solutions with patterns consisting of a single pixel.

To achieve a non-negative representation, we pixel-wise split up the derivations into a positive and negative part. For the spatial derivation we get $I_{i,d}$, with $d \in \{x+, y+, x-, y-\}$. For the temporal derivation we get $I_{i,t+}$ and $I_{i,t-}$. We apply the same procedure to the motion vector fields and get

$$\mathbf{R}_{i,d} = \sum_j \mathrm{conv}_2(\mathbf{H}_{ij,d}, \mathbf{W}_j), \quad (5)$$

with the now four dimensional activities $\mathbf{H}_{ij,d}$. We now have an explicitly non-negative representation segregated into the four orthogonal directions *right*, *up*, *left* and *down*, which leads to the non-negative BCE

$$E_{BCE,i} = \frac{1}{2}\|(\mathbf{I}_{i,x+} - \mathbf{I}_{i,x-}) \circ (\mathbf{R}_{i,x+} - \mathbf{R}_{i,x-}) + (\mathbf{I}_{i,y+} - \mathbf{I}_{i,y-}) \circ (\mathbf{R}_{i,y+} - \mathbf{R}_{i,y-}) + (\mathbf{I}_{i,t+} - \mathbf{I}_{i,t-})\|_2^2. \quad (6)$$

We achieve sparsity in the activities by penalizing the amplitude of each four dimensional activity vector $\mathbf{h}_{ij}(\mathbf{m})$ with the sparsity energy function

$$E_{H,i} = \sum_{j,\mathbf{m}} \|\mathbf{h}_{ij}(\mathbf{m})\|_2. \quad (7)$$

As pointed out in [10] the sparsity term stresses the need to normalize the basic patterns. We apply a Euclidean normalization

$$\bar{\mathbf{W}}_j = \frac{\mathbf{W}_j}{\|\mathbf{W}_j\|_2} \quad (8)$$

on the basic patterns and consider the inner derivation $\nabla_{\mathbf{W}_j}\bar{\mathbf{W}}_j(\mathbf{W}_j)$ and the chain rule for the update of $\mathbf{W}_j$ as in [10].

Our non-negative representation allows the co-existence of the reconstruction of opposing directions for each pixel, which obviously makes no sense. We thus penalize this overlap with the energy function

$$E_{R,i} = \mathbf{R}_{i,x+}^\top \mathbf{R}_{i,x-} + \mathbf{R}_{i,y+}^\top \mathbf{R}_{i,y-}. \quad (9)$$

B. Estimating the Flow Field

Given the image derivations and a set of basic patterns $\mathcal{W}$ we can now estimate the optical flow by minimizing the energy function (6) with respect to the activities $\mathbf{h}_{ij,d}(\mathbf{m})$.

Another gain from our non-negative representation is, that we can use the fast and parameter free multiplicative gradient descent method introduced by [9] to learn our activities. The resulting algorithm is as follows:

1) Initialize $\mathbf{H}_{ij,d}$ randomly,
2) Calculate $R_{i,d}$ according to eq. (5),
3) Update $\mathbf{H}_{ij,d} = \mathbf{H}_{ij,d} \circ \frac{\left(\nabla_{\mathbf{H}_{ij,d}} E\right)^-}{\left(\nabla_{\mathbf{H}_{ij,d}} E\right)^+}$,
4) Repeat from 2.) until convergence.

The gradients of the energy function (6) for $\mathbf{H}_{ij,d}$ are:

$$\left(\nabla_{\mathbf{H}_{ij,d}} E_{BCE,i}\right)^+ = \mathrm{corr}_2\left((\mathbf{I}_{i,d} \circ \mathbf{A} + \mathbf{I}_{i,\hat{d}} \circ \mathbf{B}), \mathbf{W}_j\right), \quad (10)$$

$$\left(\nabla_{\mathbf{H}_{ij,d}} E_{BCE,i}\right)^- = \mathrm{corr}_2\left((\mathbf{I}_{i,d} \circ \mathbf{B} + \mathbf{I}_{i,\hat{d}} \circ \mathbf{A}), \mathbf{W}_j\right), \quad (11)$$

with

$$\mathbf{A} = \mathbf{I}_{i,t+} + \sum_d \mathbf{I}_{i,d} \circ \mathbf{R}_{i,d}, \quad (12)$$

$$\mathbf{B} = \mathbf{I}_{i,t-} + \sum_d \mathbf{I}_{i,d} \circ \mathbf{R}_{i,\hat{d}}, \quad (13)$$

where corr_2 is the two-dimensional correlation and $\hat{d}$ describes the opposing direction to d (f.e.: $\hat{d} = x-, d = x+$). The gradient for the sparsity energy function (7) is

$$\left(\nabla_{\mathbf{H}_{ij,d}} E_{H,i}\right)^+ = \frac{\mathbf{H}_{ij,d}}{\|\mathbf{h}_{ij}(\mathbf{m})\|_2}. \quad (14)$$

The gradients for the penalization of the opposing directions (9) are

$$\left(\nabla_{\mathbf{H}_{ij,d}} E_{R,i}\right)^+ = \mathrm{corr}_2\left(\mathbf{R}_{i,d}, \mathbf{W}_j\right) \quad (15)$$

and

$$\left(\nabla_{\mathbf{W}_j} E_{R,i}\right)^+ = \sum_d \mathrm{corr}_2\left(\mathbf{R}_{i,\hat{d}}, \mathbf{H}_{ij,d}\right). \quad (16)$$

C. Learning the Basic Patterns

In related work [2], [3] the used basic patterns were pretrained on ground truth optical flow data. This approach is somehow flawed, since ground truth optical flow data is not available for a huge set of problems and we want to apply our approach to problems where other optical flow estimation algorithms have deficiencies in the task of preserving the relevant movements.

Furtunately, we can learn the basic patterns directly while estimating the optical flow. Instead of learning only the activities, for this purpose we update the basic patterns and activities alternately, using the multiplicative NMF updates, which leads to the following algorithm:

1) Initialize $\mathbf{H}_{ij,d}$ and $\mathbf{W}_j$ randomly,
2) Calculate $R_{i,d}$ according to eq. (5),
3) Update $\mathbf{H}_{ij,d} = \mathbf{H}_{ij,d} \circ \frac{\left(\nabla_{\mathbf{H}_{ij,d}} E\right)^-}{\left(\nabla_{\mathbf{H}_{ij,d}} E\right)^+}$,
4) Calculate $R_{i,d}$ according to eq. (5),
5) Update $\mathbf{W}_j = \mathbf{W}_j \circ \frac{\left(\nabla_{\mathbf{W}_j} E\right)^-}{\left(\nabla_{\mathbf{W}_j} E\right)^+}$
6) Repeat from 2.) until convergence.

Once a set of basic patterns has been learned on a training dataset, the basic patterns can be used to estimate new data. The gradients for the learning of the basic patterns are:

$$\left(\nabla_{\mathbf{W}_j} E_{BCE,i}\right)^+ = \sum_d \mathrm{corr}_2\left((\mathbf{I}_{i,d} \circ \mathbf{A} + \mathbf{I}_{i,\hat{d}} \circ \mathbf{B}), \mathbf{H}_{ij,d}\right), \quad (17)$$

$$\left(\nabla_{\mathbf{W}_j} E_{BCE,i}\right)^- = \sum_d \mathrm{corr}_2\left((\mathbf{I}_{i,d} \circ \mathbf{B} + \mathbf{I}_{i,\hat{d}} \circ \mathbf{A}), \mathbf{H}_{ij,d}\right). \quad (18)$$

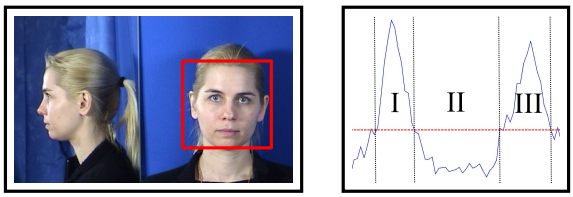

Fig. 1. Two step preprocessing: In the first step the frontal face for each frame of the video is extracted with a face detector [18]. Then the video is devided into three phases: I Onset (AU movement), II Apex (emotion is fully expressed) and III Offset (muscle relax and face returns to neutral expression). The beginning of the onset phase is automatically detected by thresholding (red line) the sum over all pixels of each differential images, shown for one sequence in the right frame (x-axis: time, y-axis: sum of absolute differences).

D. Algorithm Summary

Each incoming image pair is normalized, so that its maximum value is 1. Afterwards the spatial and temporal derivations are computed and pixel-wise split up into their positive and negative components. Next, the optical flow is estimated by minimizing the overall energy function

$$E = \sum_i \Big(E_{BCE,i} + E_{R,i} + \lambda_H E_{H,i} \Big), \qquad (19)$$

by learning the activities via gradient descent using the algorithm described above.

III. Face Movements

We applied our algorithm on videos of the MMI face expression database [14] which provides videos of humans expressing the 6 basic emotions. Each displayed emotion consists of three phases. It starts with the neutral expression, followed by the onset (phase I), where most of the movements occur. At the end of the onset phase comes the apex (phase II), where the facial expression is strongest. Finally, in the offset (phase III), the muscles relax and the face returns to its neutral expression.

A. Preprocessing

Since we are only interested in the movements of the face during the onset phase we perform a two step automated preprocessing that is depicted in figure 1. First we cut out the frontal face with the help of an automated face detector [18] and rescale it to a resolution of 128x128 pixels. Then we detect the start of the onset phase by thresholding the differential images. Note that more sophisticated systems that can robustly detect the different emotional phases exist [16]. We limit our analysis to the first ten frames of the onset phase, because that is when the main AU-related movements occur.

B. Action Units

The major gain of our algorithm is that it can robustly estimate the movement of local motion patterns. We demonstrate this for the movement of facial AU during the expression of emotions. AU are caused by facial muscles and are strongly related to 6 basic emotions described in [13]. In table I some examplary relations between emotions and AU, are depicted.

In a recent survey on emotion- and AU-recognition [15] it is stated that emotion recognition can be solved with current approaches, even though most methods neglect motion information. Contrariwise automated AU detection remains an unsolved challenge, who's solution would allow for more sophisticated face expression analysis. An automated AU detection system must solve two sub-tasks: First, extract the motion information. Then localize and relate the movement to facial parts (eyebrow, mouth, etc.). Our learning algorithm can solve the first task.

Emotion	AU	Description
Fear, Sadness, Surprise	1	Inner brow raise
Fear, Surprise	2	Outer brow raise
Anger, Sadness,	4	Brow lower
Happiness	6	Cheek raise
Anger	7	Lid tightend
Disgust	9	Nose wrinkle
Happiness	12	Lip corner puller
Disgust, Sadness	15	Lip corner depresser
Disgust	16	Lower lip depresser
Fear, Surprise	26	Jaw drop

TABLE I
SUBSET OF POSSIBLE AU AND THEIR RELATED EMOTIONS, ACCORDING TO [13].

C. Experimental Results

We now show how our motion estimation algorithm can extract the facial AU. First, the basis vectors $\mathcal{W}$ are learned on a training set with two persons showing all 6 emotions. The learning experiments are further discussed in section IV. We use the learned basis vectors to estimate the movement between each of the consecutive frames of the first ten images of the onset phase of each video. The movement of the different AU's can occur anytime during the onset phase, e.g., the narrowing of the inner eyebrows in the anger video depicted in figure 2 occurs between the third and fourth image of the onset phase. To overcome these temporal variations, we sum up the single flow fields and obtain an integrated flow field for all the movements that occur during the onset phase. The summed flow field estimated with our algorithm, the classic Horn and Schunk (HS) [5] OFE, the Lucas and Kanade (LK) [6] OFE and a state-of-the-art segmentation-based OFE method by Sun et al. [7] for one person performing all 6 emotions are shown in figure 2.[1]

We focus the discussion on whether the AU movement was preserved, if the movement can be locally assigned to a facial part and how robustly the algorithms work. The HS OFE can capture most of the movements, but looses all local

[1]For the HS and [7] algorithm we used the publicly available implementation described in [7]. For the HS method we choose the carbonier error function (method 'classic-c'). We performed manual parameter optimization for the regularization parameter and found no qualitative difference in the resulting flow fields. For the displayed flow the default parameters were used.

shape relations due to its strong dependency on the spatial regularization term. The LK OFE can preserve most of the local relations, but has a lot of miss-estimations and is the least robust method.[2] The highly sophisticated OFE by Sun et al. [7] is capable of preserving an impressive amount of detail due to its local segmentation. Unfortunately similar to the HS algorithm it tends to overregularize in low contrast areas as e.g. the cheeks and thus looses most of the local relations.

In contrary to the above mentioned methods, our algorithm is capable of extracting the main AU related movements and keeps the local shape relations robustly. E.g. the subtle narrowing of the eyebrows in the anger sequence is detected (small red and blue patterns) as well as AU 1 (inner brow raise) during the fear sequence (small purple patterns).

IV. Learning Motion Patterns

The purpose of each basic pattern $\mathbf{W}_j$ is to describe the local neighbourhood for which the assumption that on a local scale pixels of the same rigid object move coherently holds. Thus each $\mathbf{W}_j$ describes the area where the pixels have a similar tanslational movement. $\mathbf{W}_j$ functions as a weighted mask for the corresponding activity $h_{ij,d}(\mathbf{m})$ which describes the movement direction of the basic pattern shifted to the specific position $\mathbf{m}$.

We now want to learn this local neighbourhood with our unsupervised learning algorithm. The two free parameters are the number of basic patterns J and the sparsity parameter λ_H. In the following we discuss the influence of those two parameters.

A. Influence of Sparsity

The sparsity parameter λ_H regulates the influence of the sparsity energy function E_H (7) and thus the amount of activity used for the construction of the optical flow field. There are two main reasons for enforcing sparsity for our reconstruction. First, with sparsity the learned patterns tend to be spatially more extended, thus the trivial solution, where a basic pattern consist of just a single pixel, is averted. Second, since as few activities as possible are used for the reconstruction, it is more likely that each basic pattern has to describe its part of the flow field on its own, without overlaying it with other basic patterns. So each basic pattern will become more specific.

In summary, the influence of the two energy functions E_{BCE} (4) and E_H (7) can be understood as follows: On the one hand E_{BCE} will force the reconstruction $\mathbf{R}$ to fullfil the BCE as well as possible and will thus favor arbitrary basic patterns. On the other hand E_H enforces the basic patterns to be less arbitrary and more specifc. The weighting parameter λ_H has to be choosen in a way that a desirable mixture of both objectives is aquired. The learned basic patterns should be both specific but generalisable.

[2]We had to discard outliers and inserted a bias to zero movement by adding a constant noise image to all images of a sequence, to get at least some reasonable results.

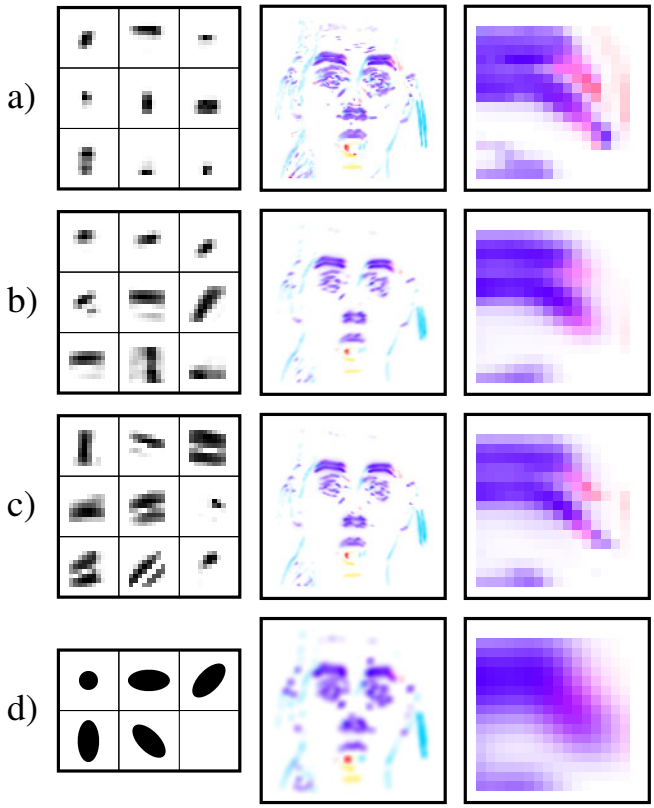

Fig. 3. Learned basic patterns: The first column shows up to 9 example basic patterns learned for different parameter settings. The second column displays a flow field for a video of the surprise emotion, that is estimated with the basic patterns of the same row and the parameter $\lambda_H = 0.00001$. The third column shows the zoomed flow field of the right eyebrow movement for the same flow field. The different parameters are: a) $\lambda_H = 0.001$, b) $\lambda_H = 0.01$, c) $\lambda_P = 0.01$ and $\lambda_H = 0.01$. Row d) shows a set of handcrafted basic patterns.

In figure 3 a subset of the learned basic patterns for the two parameters $\lambda_H = 0.001$ (figure 3 a)) and $\lambda_H = 0.01$ (figure 3 b)) are shown together with their corresponding reconstruction. The basic patterns for smaller λ_H are less spatially extended and can therefore achieve a better description of detailed regions as depicted in the zoomed reconstruction (right column in figure 3). Contrariwise the reconstruction (middle column in figure 3) for the basic patterns learned with higher sparsity is less prone to noise and focuses on the more relevant motions.

The basic patterns learned with $\lambda_H << 0.001$ converge to the trivial solution (one pixel basic patterns), while the activities for $\lambda_H >> 0.01$ converge to zero, which is another trivial solution (no flow field). We found that parameters $0.001 < \lambda_H < 0.01$ lead to good basic patterns for a variety of different videos like facial movements, gestures, full body movements, animal movements etc.[3]

Once a set of basic patterns $\mathcal{W}$ is learned, the sparsity parameter becomes less relevant, therefore we set $\lambda_H =$

[3]Note that we have normalized each incoming image between 0 and 1 before calculating the gradients. Other normalizations might lead to different λ_H values.

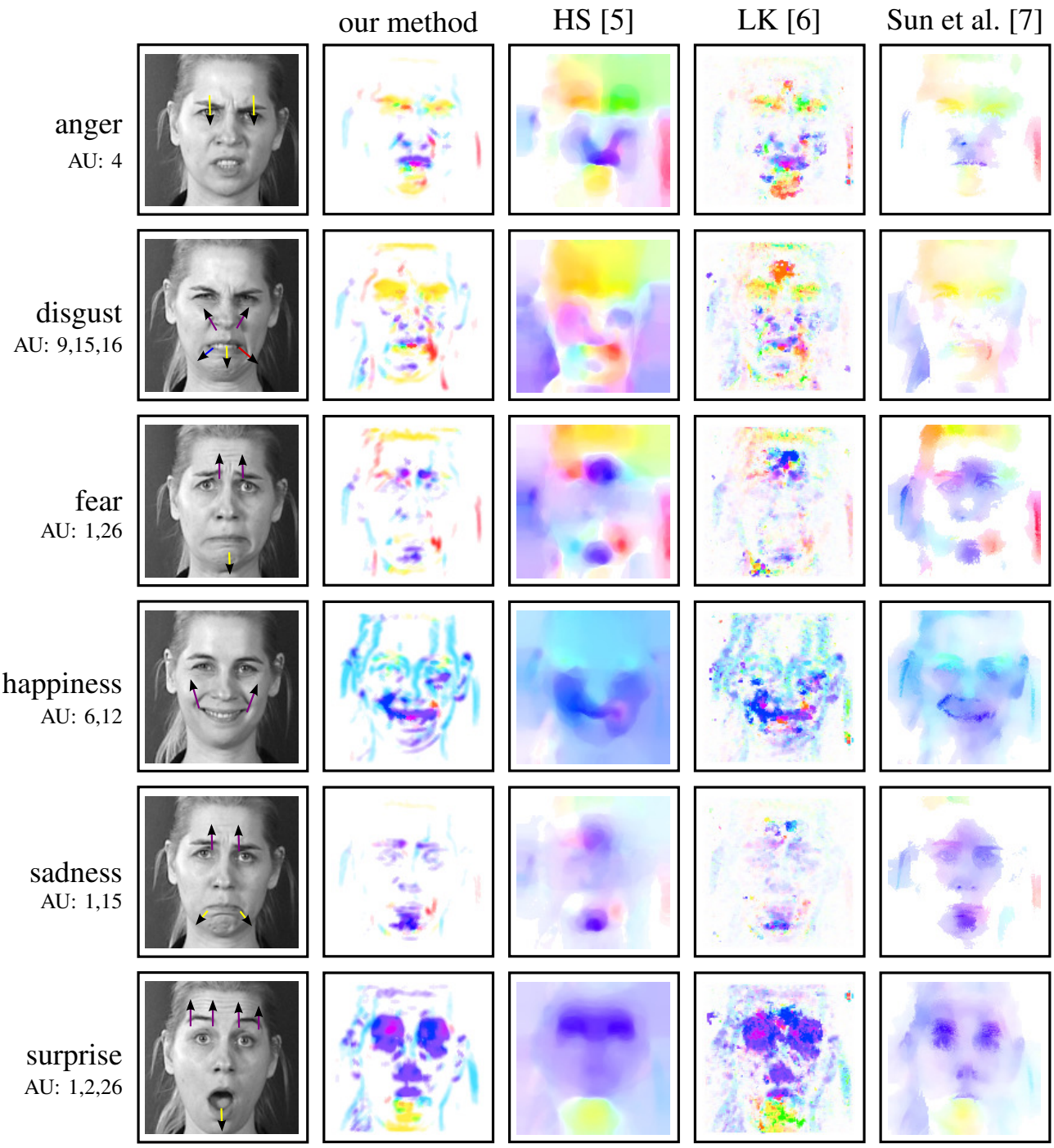

Fig. 2. Method comparison: The first column presents the first image of the apex phase (emotion fully expressed) for each emotion. In addition some of the emotion-related AU are drawn onto the image (arrows). The other columns show the color-coded integrated flow fields for the four different algorithms, i.e. the integrated flow fields of each consecutive image pair of the onset phase (face movement from neutral to apex). The color code is taken from [7] and works as follows: Each color indicates a direction (red = right, purple = up, blue = left, yellow = down) while the velocity is encoded in the saturation.

0.00001 for all experiments with pre-trained basic patterns.

B. Number and Size of Basic Patterns

The influence of the number of basic patterns J is basically as follows. The higher J, the more detailed patterns will be learned as a result. If only one basic pattern is learned it looks like a gaussian blob. We were able to learn up to 25 different basic patterns. If J is choosen above 25, the resulting patterns become redundant. For low sparsity values this effect already starts for $J > 9$.

We limit the maximal size of the basic patterns to be rather small (8x8 pixels). The main reason is, that the assumption that pixels in a local neighbourhood have the identical translational movement is only fullfilled for a small local

area. This is in good accordance to our experiments, where we observed that larger basic patterns lead to erroneous motion estimations. Another argument for local patterns is, that larger basic patterns get more situation specific and are therefore less able to generalize to new video data.

C. Handcrafted Basic Patterns

To demonstrate the effectiveness of learning the basic patterns we compare them to a set of handcrafted basic patterns. Inspired by the shape of our learned patterns we designed one round (radius of 4 pixels) and four rotated ellipsoid-shaped patterns (8x4 pixels) that are depicted in figure 3 d). While the main motion information is optained, most of the detailed information that the learned basic patterns can preserve is lost.

D. Enforced Parts-basedness

We can further enforce a more parts-based decomposition by directly penalizing overlaps of the partial reconstructions $\mathbf{R}_{ij\mathbf{m},d}$ with the additional energy function

$$E_{P,i} = \frac{1}{2} \sum_{d,j,\mathbf{m}} \Big(\mathbf{R}_{ij\mathbf{m},d} \Big)^\top \Big(\sum_{k,\mathbf{n}} \mathbf{R}_{ik\mathbf{n},d} - \mathbf{R}_{ij\mathbf{m},d} \Big), \quad (20)$$

that was introduced in our previous work [1]. The gradients are

$$\Big(\nabla_{\mathbf{H}_{ij}} E_{P,i} \Big)^+ = \mathrm{corr}_2(\mathbf{R}_{i,d}, \mathbf{W}_j), \quad (21)$$

$$\Big(\nabla_{\mathbf{H}_{ij}} E_{P,i} \Big)^- = \mathbf{H}_{ij} \mathbf{W}_j^\top \mathbf{W}_j, \quad (22)$$

for the activities and

$$\Big(\nabla_{\mathbf{W}_j} E_{P,i} \Big)^+ = \sum_d \mathrm{corr}_2(\mathbf{R}_{i,d}, \mathbf{H}_{ij,d}), \quad (23)$$

$$\Big(\nabla_{\mathbf{W}_j} E_{P,i} \Big)^- = \mathbf{W}_j \sum_d \mathbf{H}_{ij,d}^\top \mathbf{H}_{ij,d}, \quad (24)$$

for the basic patterns. The new total energy function is now

$$E = \sum_i \Big(E_{BCE,i} + E_{R,i} + \lambda_H E_{H,i} + \lambda_P E_{P,i} \Big), \quad (25)$$

with the new parameter λ_P that weights the influence of the penality function (20). The influence of the new energy function for the parameters $\lambda_P = 0.01$ and $\lambda_H = 0.01$ is depicted in figure 3 c). The basic patterns tend to represent prototypical parts of the flow field and are thus very specific and less generalisable. The reconstruction combines the benefits of the small and large sparsity values, because most of the motion details are preserved and in addition the flow field focuses on the relevant motion patterns and discards noisy single pixel movements.

E. Parameter Settings

The parameter discussion can be summerized as follows: The maximum size of the basic patterns should be set to partly cover the AU (in our case 8x8 pixels). The number of basic patterns J can be choosen between 9-25, depending on the required level of detail, which can further be enhanced by adding the parts-based energy functional (20), with the parameter setting: $0.05 > \lambda_P > 0.01$. The sparsity parameter for learning should be set between $0.01 > \lambda_H > 0.001$, again depending on the required level of detail.

For our experiments we set $J = 9$, $\lambda_P = 0.01$, $\lambda_H = 0.01$ for learning and $\lambda_H = 0.00001$ for detection of the movement with the learned basic patterns.

V. Emotion Recognition

We want to further demonstrate the benefit of our motion features using a simple emotion recognition setup. We choose a subset of the MMI face expression database [14], consisting of ten persons for each of the 6 basic emotions. For each of the videos the flow field is calculated using a set of basic patterns that was pretrained on another subset of the database, showing other persons in front of a different background. The goal is now to evaluate if it is possible to distinguish two emotions based on motion features alone. We perform leave-one-out experiments for all possible emotion combinations.

A. Feature and Emotion Classification

For each video we extract one feature vector out of the automatically detected onset phase. Each feature vector is calculated as follows: First the nine onset flow fields are calculated and summed up to the integrated flow field which contains all the movement that occurs during the onset phase. Then the center area of the flow field is splitted into six non-overlapping square cells as depicted in figure 4. The two upper cells each roughly cover the eyes and eyebrows, the middle cells the nose and upper lips and the two lower cells cover each side of the lower lips and the jaw. In each cell all values for each of the four non-negative directions (right, up, left, down) of the flow fields are summed up and stored in the feature vector, which thus consists of 24 entries.

For the classification we use a *support vector machine* (SVM) [8] with a radial basis function kernel. The kernel parameter was trained using 3-fold cross-validation. The entire setup is shown in figure 4.

B. Results

The detailed results for the flow fields calculated with our algorithm are shown in table II. The results vary strongly between the different emotion pairs. While some pairs have recognition rates of 95% and are thus easy distinguishable with motion information, others, especially the emotion pair fear-surprise are not well separable using motion information only.

The mean recognition rate over all emotion pairs for our method and the other three discussed OFE methods are shown in table III. While our method achieves the highest

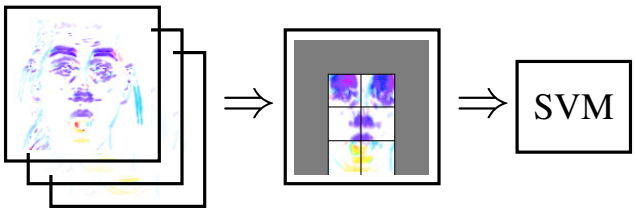

Fig. 4. Classification framework: All the optical flow fields of the onset phase are summed up. Then the lower central region is divided into six square cells. For each cell the values of the four non-negative directions are summed up. The 24 dimensional feature vector for each video is then classified using a support vector machine.

anger-disgust	anger-fear	anger-happiness
90%	95%	95%
anger-surprise	disgust-fear	disgust-happiness
95%	65%	95%
disgust-surprise	fear-happiness	fear-sadness
85%	70%	80%
happiness-sadness	happiness-surprise	sadness-surprise
80%	80%	90%
anger-sadness	disgust-sadness	fear-surprise
65%	80%	40%

TABLE II
DETAILED RESULTS FOR OUR METHOD.

recognition rate, the difference to the two regularization-based approaches (Horn and Schnunk and Sun et al.) is rather marginal. The reason for this is that during the pooling in each cell most of the local information is lost, so that the advantages of precisely localized motion areas for the AU as extracted by our method cannot be fully exploited. Note that no preprocessing except the face and onset detection is performed on the original videos. Future work should combine the obtained motion information with a fine and robust spatial segmentation to benefit more from the local character of the extracted motion patterns.

our method	Sun et al. [7]	HS [5]	LK [6]
81%	80%	79%	77%

TABLE III
MEAN CLASSIFICATION RESULTS FOR DIFFERENT MOTION ESTIMATION METHODS.

VI. SUMMARY & CONCLUSION

We present a novel unsupervised learning algorithm based on a generative model that can learn sets of basic patterns which describe small local regions of coherent motion to reconstruct optical flow fields. These basic patterns enable us to robustly extract motion information while preserving local shapes of small motion patterns such as facial AU.

To show that a superposition of learned basic motion patterns is a quite flexible and detailed representation of visual motion, we further used the basic patterns learned on the facial data to estimate the motion for different kinds of natural movements as shown in figure 5. As can be seen



Fig. 5. Other movements: The three pictures show human and animal movement estimated with the basic patterns learned on the face dataset and the same parameters as for the face movements.

the basic patterns that are quite suitable for face expression recognition are also capable to reconstruct for example human body motions and animal movements. This implies using a limited number of basic motion patterns we are able to generate any kind of natural movements independently on what kind of object is moving.

REFERENCES

[1] T. Guthier, J. Eggert and V. Willert, "Unsupervised learning of motion patterns", *European Symposium on Artificial Neural Networks* (ESANN), 2012.

[2] K. Jia, X. Wang and X. Tang, "Optical flow estimation using learned sparse model", *IEEE Conf. on Computer Vision* (ICCV), pp. 2391-2398, 2011.

[3] C. Nieuwenhuis, D. Kondermann and C. S. Garbe, "Complex motion models for simple optical flow estimation", *Proc. of the 32nd DAGM conf. on Pattern recognition*, 2010.

[4] C. F. Cadieu and B. A. Olshausen, "Learning intermediate-level representations of form and motion from natural movies", *Neural computation*, 24.4, pp. 827-866, 2012.

[5] B. Horn and B. Schunck, "Determining optical flow", *Artificial Intelligence*, 16, pp. 185-203, 1981.

[6] B. Lucas and I. Kanade, "An iterative image registration technique with an application to stereo vision", *DARPA Image Understanding Workshop*, pp. 121-130, 1981.

[7] D. Sun, S. Roth and M.J. Black, "Secrets of optical flow estimation and their principles", *IEEE Conf. on Computer Vision and Pattern Recognition* (CVPR), pp. 2432-2439, 2010.

[8] C.C. Chang and C.J. Lin., "LIBSVM : a library for support vector machines", *ACM Transactions on Intelligent Systems and Technology*, 2011.

[9] D.D. Lee and S. Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature*, vol. 401, pp. 788-791, 1999.

[10] J. Eggert and E. Koerner, "Sparse coding and NMF", *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), vol. 4, pp. 2529-2533, 2004.

[11] J. Eggert, H. Wersing and E. Koerner, "Transformation-invariant representation and NMF", *IEEE Int. Joint Conf. on Neural Networks* (IJCNN), pp. 2535-2539 vol. 4, 2004.

[12] B. Olshausen and D.J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images", *Nature*, vol. 381, pp. 607-609, 1996.

[13] P. Ekman, W.V. Friesen and J.C. Hager, "Facial action coding system", UT : A Human Face, 2002.

[14] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database", *The Workshop Programme*, 2010.

[15] M. Valstar, M. Mehu, B. Jiang, M. Pantic and Klaus Scherer, "Meta-Analysis of the First Facial Expression Recognition Challenge", *IEEE Trans. on Systems, Man, and Cybernetics-B*, vol. 42(4), pp. 966-979, 2012.

[16] M. Valstar and M. Pantic, "Fully Automatic Recognition of the Temporal Phases of Facial Actions", *IEEE Trans. on Systems, Man, and Cybernetics-B*, 42(1), pp. 28-43, 2012.

[17] J. Haag and A. Borst, "Neural mechanism underlying complex receptive field properties of motion-sensitive interneurons", *Nature Neuroscience*, vol. 7, pp. 628-634, 2004.

[18] P. Viola and M. J. Jones, "Robust real-time face detection", *Int. journal of computer vision*, 57.2, pp. 137-154, 2004.