Investigation of Small Group Social Interactions Using Deep Visual Activity-Based Nonverbal Features

Cigdem Beyan[1], Muhammad Shahid[1,2], Vittorio Murino[1,3]

[1] *Pattern Analysis and Computer Vision, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genoa, Italy*
[2] *Elect., Electron. and Telecom. Eng. and Naval Arch. Department, University of Genoa, 16126, Genoa, Italy*
[3] *Department of Computer Science, University of Verona, Ca' Vignal 2, Strada Le Grazie 15, 37134, Verona, Italy*
{Cigdem.Beyan, Shahid.Muhammad, Vittorio.Murino}@iit.it

ISTITUTO ITALIANO DI TECNOLOGIA

Contact:
Cigdem.Beyan@iit.it
https://www.iit.it/people/cigdem-beyan

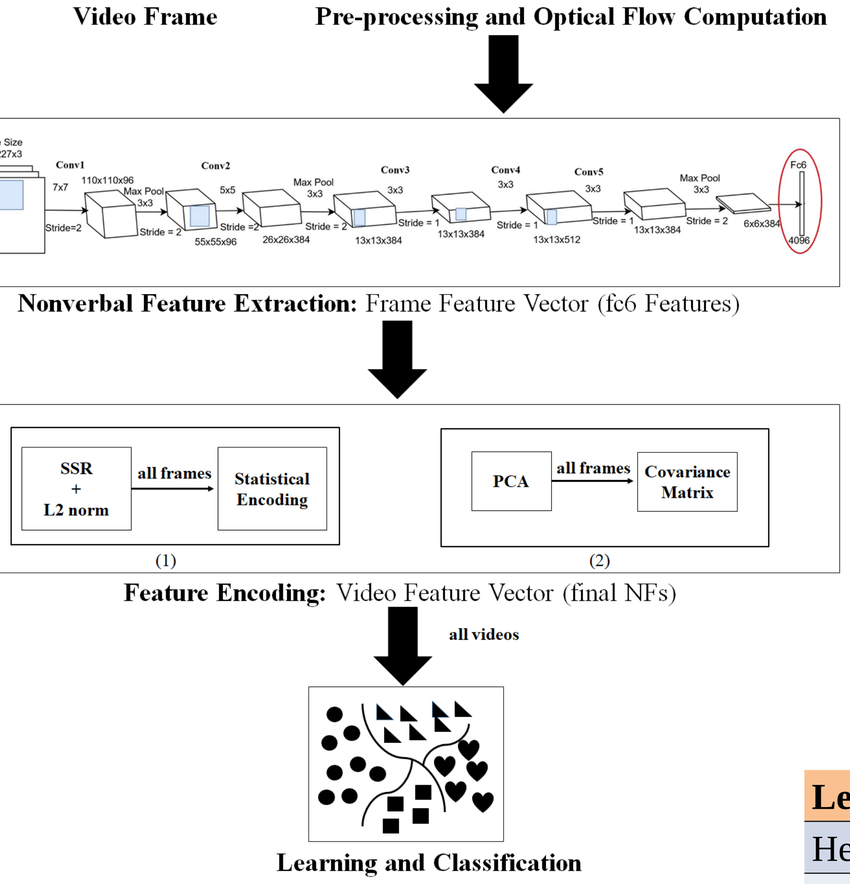
MOTIVATION

- Automatic understanding of **small group social interactions**:
  - Popular research topic in social computing (dominance estimation, role recognition, emergent leadership, personality traits classification, etc.).
- **Nonverbal Features (NFs):**
  - Among core research problems in social psychology,
  - High amount of information,
  - Eye gaze, head/body activity, speaking activity, energy, pitch, etc.
  - Audio-based, video-based, and audio-visual.
- **Visual Activity (VisualAct):**
  - Important cue to understand various interactions,
  - Implemented in terms of: head/body activity, weighted motion energy image, body pose,
  - Usually not as good as other NFs.

CONTRIBUTIONS

- **Novel VisualAct-based NFs**:
  - first time: **Convolutional Neural Network** is used,
  - showing **(significantly) improved results** as compared to the state of the art VisualAct-based NFs and when they are combined with other NFs/modalities.
- First time: **Covariance** is used to encode features extracted from a CNN model:
  - performing as well as other feature encoding method, but resulting in **much low-dimensional feature vectors.**

PROPOSED METHOD



- Applied to 3 problems**:**
  - Emergent leader detection [1]
  - Prediction of leadership styles [1]
  - High/low extraversion classification [2]
- **Pre-processing:** Image cropping; only participant and a small amount of background remain.
- **Optical flow image computation** [3]**:** An RGB image constructed from *x*, *y* flow values and the flow magnitude. Zero mean normalized.
- **NF Extraction:** a pre-trained CNN model [4], no fine-tuning.
  - Values in fc6 layer used.
- **Feature Encoding:** fc6 features are for each flow image (*frame feature vector*) and combined into a single feature vector i.e. *video feature vector* using:
  - **Statistical Encoding (STAT)** [5]: Normalization w/signed square root and L2, extract mean, variance, min., max., etc., then z-score normalization to obtain final NFs.
  - **Covariance Encoding (CM):** Principal component analysis, then covariance matrix calculation from Riemannian manifold to Euclidean space. *Final NFs:* all entries on and above (below) the diagonal of the covariance matrix.
- **Learning and Classification:** Support Vector Machine (SVM) and Localized Multiple Kernel Learning (LMKL) [6].

Emergent Leader (EL) Detection	
Head Activity- SVM [7]	0.48
Body Activity- SVM [7]	0.46
Head/Body Activity- LMKL [7]	0.59
Body Pose- LMKL [7]	0.64
Speaking/Head/Body Activity- LMKL [7]	0.74
Proposed Method- SVM	0.53
Proposed Method- LMKL	**0.76**

Leadership Style (LS) Prediction	
Head Activity- SVM [8]	0.40
Body Activity- SVM [8]	0.41
Gaze/Head/Body Activity- LMKL [8]	0.69
Gaze/Speaking/Head/Body Activity- LMKL [8]	0.69
Proposed Method- SVM	0.46
Proposed Method- LMKL	**0.71**

High/Low Extraversion (Ext.) Classification	
[9]	0.67
[10]	0.70
Head/Body Activity [11]	0.71
Weighted motion energy image [11]	0.66
[11]	**0.77**
Proposed Method- SVM	0.64
Proposed Method- LMKL	0.72

- **EL Detection/ LS prediction,** proposed method with LMKL
  - Better than all others, even as compared to methods using multi-modal cues,
  - Significantly better results (p-value<0.05) than other VisualAct-based NFs.
- **Ext. Classification**
  - Proposed method is video-only and with LMKL it was better than many other SOA. SOA uses various multi-modal NFs and less simple techniques.
- **STAT/CM:** Which is better? Depends on application.
  - CM resulted in 10 NFs while STAT resulted in 20480 NFs.
- **Fine-tuning** results were worse than any baseline and the proposed method.

	STAT-SVM	CM-SVM	STAT-LMKL	CM-LMKL	Fine Tune
EL Detection	**0.53**	**0.53**	**0.76**	0.72	0.40
LS Prediction	0.40	**0.46**	0.61	**0.71**	0.38
Ext. Classification	**0.64**	**0.64**	**0.72**	0.67	0.56

References:
[1] Beyan et al., *Detecting Emergent Leader in a Meeting Environment Using Nonverbal Visual Features Only*, ACM ICMI, 2016.
[2] Sanchez-Cortes et al., *A Nonverbal Behavior Approach to Identify Emergent Leaders in Small Groups*, IEEE Trans. Multimedia, 2012.
[3] Thomas Brox et al., *High accuracy optical flow estimation based on a theory for warping*, ECCV, 2004.
[4] Donahue et al., *Long-term Recurrent Convolutional Networks for Visual Recognition and Description*. CVPR, 2015.
[5] Bargal et al., *Emotion Recognition in the Wild from Videos using Images*, ACM ICMI, 2016.
[6] M. Gonen et al., *Localized Multiple Kernel Learning*, ICML, 2008.
[7] Beyan et al., *Moving as a Leader: Detecting Emergent Leadership in Small Groups using Body Pose*, ACM Multimedia, 2017.
[8] Beyan et al., Prediction of the Leadership Style of an Emergent Leader Using Audio and Visual Nonverbal Features, IEEE Trans. Multimedia., 2018.
[9] Aran and Gatica-Perez, *One of a kind: inferring personality impressions in meetings*, ACM ICMI, 2013.
[10] Okada et al., *Personality Trait Classification via Co-Occurrent Multiparty Multimodal Event Discovery,* ACM ICMI, 2015.
[11] Kindiroglu et al., *Multi-domain and multitask prediction of extraversion and leadership from meeting videos,* EURASIP Journal on Image and Video Processing, 2017.