Multimodal Visual Concept Learning with Weakly Supervised Techniques

Giorgos Bouritsas | Petros Koutras | Athanasia Zlatintsi | Petros Maragos
School of E.C.E., National Technical University of Athens, Greece
http://cvsp.cs.ntua.gr/

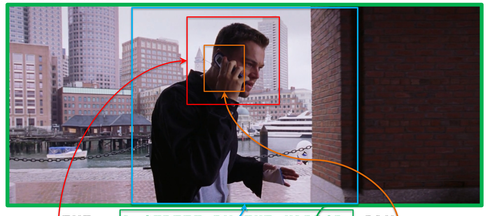
NATIONAL TECHNICAL UNIVERSITY OF ATHENS

CVPR 2018 SALT LAKE CITY • JUNE 18-22

Goal

- Detect and recognize visual concepts in videos in a weakly supervised manner, mining their labels from an accompanying descriptive text.
- **Visual Concepts***: Spatio-temporally localized video segments that carry a specific structure in the visual domain.*
- **Examples:**

1. Faces
2. Actions
3. Scenes
4. Objects



- **Motivation:**

Why Natural Language?
- Rich semantics – interpretable – easy to extract.

Why Weak Supervision?
- Reduce the time-consuming and costly procedure of manual annotation.

a) Achieve recognition in data annotated sparsely/imprecisely.

b) Collect new data to train fully supervised models.

Overview

- **Challenges:**
  - Spatio-Temporal ambiguity: absence of specific spatio-temporal correspondence between visual and textual elements.
  - Semantic ambiguity: Words/sentences may have various different meanings.
- **Contributions:**
  1. We introduce two novel weakly supervised techniques that extend the Multiple Instance Learning (MIL) framework.
     - *Fuzzy Sets Multiple Instance Learning (FSMIL)* → Spatio -Temporal ambiguity.
     - *Probabilistic Label Multiple Instance Learning (PLMIL)* → Semantic ambiguity.
  2. We propose an unsupervised, semantic similarity based method to extract weak labels from complex textual semantics.
- **Approach:**
  - Unidirectional Model: Text → Video | Extension: Bidirectional: Text ↔ Video
  - Learning method:

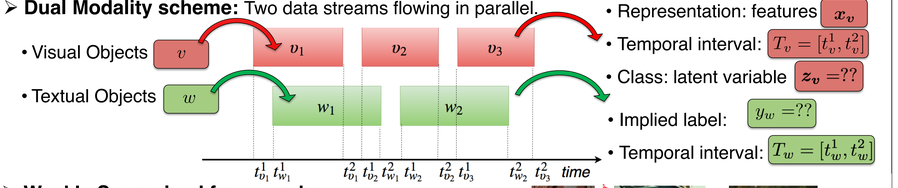
Discriminative Clustering (DIFFRAC [1]) + FSMIL & PLMIL

Convex Quadratic program *s.t* Linear Constraints

- **Evaluation:**
  - We outperform state-of-the-art methods (Bojanowski et al. 2013 [3], Miech et al. 2017[4]) on the **COGNIMUSE** dataset [2] on the tasks of face and action recognition.

Proposed Approach

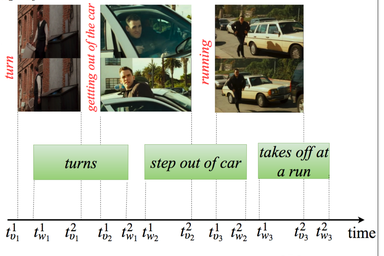
- **Dual Modality scheme:** Two data streams flowing in parallel.



- **Weakly Supervised frameworks:**
  - **Fuzzy Sets MIL (FSMIL):** Fuzzy bags of Multiple Instances.

$$\mathcal{V}_w = \{(v, \mu_w(v)) \mid v \in \mathcal{V}, \mu_w(v) = g(\frac{|T_w \cap T_v|}{|T_v|})\}$$

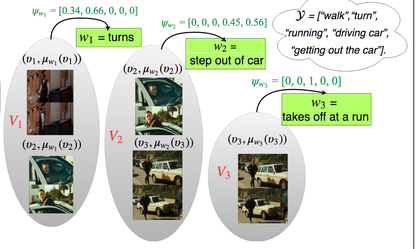
Visual objects overlapping with the textual one — *Membership grade*



  - **Probabilistic Label MIL (PLMIL):**

Each bag is assigned a probabilistic label. $\psi_w(y) = \mathbb{P}[y_w = y|w]$

Unsupervised estimation via semantic similarity. $\psi_w(y) = s_{wy} / \sum_{\ell \in \mathcal{Y}} s_{w\ell}$



- **Discriminative clustering model (DIFFRAC [1]):**
  - Ridge regression with linear classifier $\mathrm{f}(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{\omega} + \boldsymbol{b}$

$$\min_{\boldsymbol{Z},\boldsymbol{\omega},\boldsymbol{b}} \frac{1}{2V}\|\boldsymbol{Z} - \boldsymbol{X}^T\boldsymbol{\omega} - \mathbf{1}_V\boldsymbol{b}\|_F^2 + \frac{\lambda}{2}Tr(\boldsymbol{\omega}^T\boldsymbol{\omega})$$

Closed form w.r.t classifier

- **Our approach:**

Loss + Regularizer

$\mathcal{Y}_w = \{y \mid \psi_w(y) \neq 0\}$

Weighted bag members – FSMIL

$$\min_{\boldsymbol{Z},\boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{A}(\boldsymbol{X},\lambda)) + \kappa \sum_{w\in\mathcal{W}}\sum_{y\in\mathcal{Y}_w} f(\psi_w(y))\xi_{wy}^2,$$
$$s.t \quad \forall w \in \mathcal{W}, \forall y \in \mathcal{Y}_w, \sum_{v\in\mathcal{V}_w} z_{vy}\mu_w(v) \geq 1 - \xi_{wy}$$

Relaxing the constraints

Relaxation factor – PLMIL

- **Bi-directional model:**

Text→Video

Alternating optimization

Video→Text

$$\min_{\boldsymbol{Z},\boldsymbol{T},\boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{A}(\boldsymbol{X}_v,\lambda_v)) + \kappa_v \sum_{w\in\mathcal{W}}\sum_{y\in\mathcal{Y}_w} f(\psi_w(y))\xi_{wy}^2 +$$
$$Tr(\boldsymbol{T}\boldsymbol{T}^T\boldsymbol{A}(\boldsymbol{X}_w,\lambda_w)) + \kappa_w \sum_{v\in\mathcal{V}}\sum_{y\in\mathcal{Y}_v} f(\psi_v(y))\xi_{vy}^2$$

$$\psi_v(y) = z_{vy}\ ,\ \psi_w(y) = t_{wy}\ ,\ \mu_w(v) = \mu_v(w)$$

$$\forall w \in \mathcal{W}\ ,\ \forall y \in \mathcal{Y}_w \quad \sum_{v\in\mathcal{V}_w} z_{vy}\mu_w(v) \geq 1 - \xi_{wy}$$

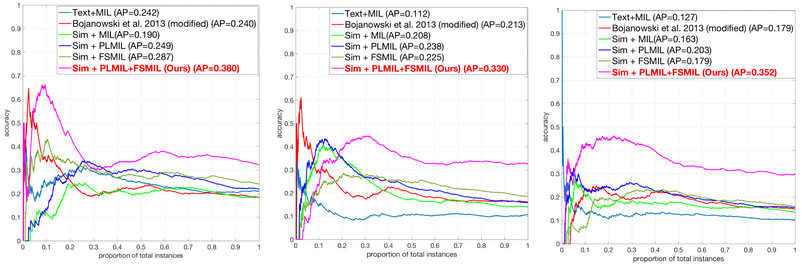
$$\forall v \in \mathcal{V}\ ,\ \forall y \in \mathcal{Y}_v \quad \sum_{w\in\mathcal{W}_v} t_{wy}\mu_v(w) \geq 1 - \xi_{vy}$$

Results

- **COGNIMUSE Dataset:** 5 movies + scripts: Dev Set (***DEP, LOR***) & Test Set (***BMI, CRA, GLA***).
- **Concept Detection and Feature Extraction:**
  - **Faces:** Detection & Tracking & Alignment: [3] Representation: VGG Kernel: min-min RBF
  - **Actions:** Detection: Manual Representation: C3D Kernel: Linear
- **Label Mining from the Text:**
  - **Faces:** cast list + string matching
  - **Actions:** short sentences + sentence similarity + similarity threshold θ
  - **Temporal intervals:** script to subtitle alignment (DTW)
- **Face Recognition:**
  - Bojanowski et al. 2013 [3]: treats both ambiguities with hard constraints (MIL).
  - Miech et al. 2017 [4]: extension of [3]. Extra constraint for background concepts.

Set	Development			Test				All
	DEP	LOR	MAP	BMI	CRA	GLA	MAP	MAP
Text+MIL	0.433	0.656	0.544	0.551	0.434	0.437	0.474	0.502
SIFT+MIL [3]	0.630	0.879	0.755	0.724	0.644	0.681	0.683	0.711
SIFT+FSMIL	0.693	0.881	0.787	0.770	0.691	0.746	0.736	0.756
VGG+MIL	0.834	**0.954**	0.894	0.825	0.696	0.830	0.784	0.828
VGG+FSMIL (Ours)	**0.864**	0.952	**0.908**	**0.857**	**0.731**	**0.901**	**0.830**	**0.861**
[4]+VGG: fg	0.788	0.898	0.843	0.666	0.479	0.577	0.574	0.682
[4]+VGG+FSMIL: fg	**0.810**	**0.913**	**0.862**	**0.696**	**0.505**	**0.651**	**0.617**	**0.715**
[4]+VGG: bg	**0.185**	**0.189**	**0.187**	**0.304**	0.047	**0.052**	0.134	0.155
[4]+VGG+FSMIL: bg	0.184	**0.189**	**0.187**	0.269	**0.278**	0.038	**0.195**	**0.192**

- **Action Recognition:** mean per sample accuracy curves for 6,8 & 10 action classes.



- **Code and data:** http://cvsp.cs.ntua.gr/research/multimodal_weakly_supervised_learning/

References

[1] F. R. Bach and Z. Harchaoui. "DIFFRAC: a discriminative and flexible framework for clustering". NIPS, 2008.
[2] A. Zlatintsi et al. "COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization." EURASIP, 2017.
[3] P. Bojanowski et al. "Finding actors and actions in movies." ICCV, 2013.
[4] A. Miech et al. "Learning from video and text via large-scale discriminative clustering." ICCV, 2017.