This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

# Multimodal Visual Concept Learning with Weakly Supervised Techniques

Giorgos Bouritsas, Petros Koutras, Athanasia Zlatintsi and Petros Maragos
School of E.C.E., National Technical University of Athens, Greece
gbouritsas@gmail.com, {pkoutras, nzlat, maragos}@cs.ntua.gr

## Abstract

*Despite the availability of a huge amount of video data accompanied by descriptive texts, it is not always easy to exploit the information contained in natural language in order to automatically recognize video concepts. Towards this goal, in this paper we use textual cues as means of supervision, introducing two weakly supervised techniques that extend the Multiple Instance Learning (MIL) framework: the Fuzzy Sets Multiple Instance Learning (FSMIL) and the Probabilistic Labels Multiple Instance Learning (PLMIL). The former encodes the spatio-temporal imprecision of the linguistic descriptions with Fuzzy Sets, while the latter models different interpretations of each description's semantics with Probabilistic Labels, both formulated through a convex optimization algorithm. In addition, we provide a novel technique to extract weak labels in the presence of complex semantics, that consists of semantic similarity computations. We evaluate our methods on two distinct problems, namely face and action recognition, in the challenging and realistic setting of movies accompanied by their screenplays, contained in the COGNIMUSE database. We show that, on both tasks, our method considerably outperforms a state-of-the-art weakly supervised approach, as well as other baselines.*

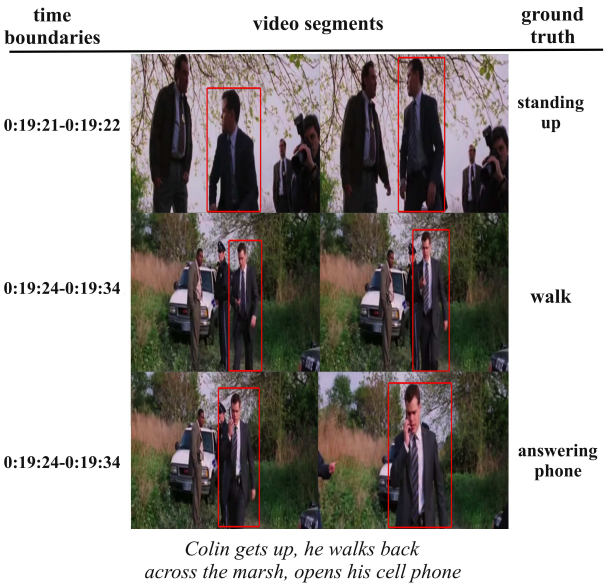



Figure 1: Example of a video segment described by the text shown below the pictures. During the time interval [0:19:21 - 0:19:34] three actions take place ("standing up", "walk", "answering phone") performed by the same person (Colin). The corresponding text mentions the actions as "gets up", "walks back" and "opens his cell phone", respectively.

## 1. Introduction

Automatic video understanding has become one of the most essential and demanding challenges and research directions. The problems that span from this field, such as activity recognition, saliency and scene analysis, comprise detecting events and extracting high level semantics in realistic video sequences. So far, the majority of the methods designed for these tasks deal with visual data ignoring the presence of other modalities, such as text and sound. Nonetheless, the exploitation of the information they provide can lead to better understanding of the underlying semantics. In addition, most of these techniques are fully supervised and are trained on diverse and usually large-scale datasets. Recently, in an attempt to avoid the significant cost of manual annotation, there has been an increasing interest in exploring learning techniques that reduce human intervention.

Motivated by the above, in this paper we approach video understanding multimodally, where our goal is to recognize visual concepts mining their labels from an accompanying descriptive document. Visual concepts could be loosely defined as spatio-temporally localized video segments that carry a specific structure in the visual domain, which allows them to be classified in various categories. Some specific examples are human faces, actions, scenes, objects etc. The main reason for using text as a complementary modality is the convenience that natural language provides in expressing semantics. Nowadays, there is a plethora of video data with natural language descriptions,

*i.e*. videos on YouTube [30, 31, 45], TV broadcasts including captions [9], videos from parliament or court sessions accompanied by transcripts [24] and TV series or movies accompanied by their subtitles, scripts, or audio descriptions [7, 11, 18, 19, 23, 38, 49]. The last category has recently gathered much interest, mainly because of the descriptiveness of these texts and the realistic nature of the visual data. Inspired by such work, we apply our algorithms to movies accompanied by their scripts. In Figure 1 we illustrate an example that was extracted from a movie, in which different instances of the action visual concept are described by an accompanying text segment.

Towards this goal, we use a unidirectional model, where information flows from text to video data. This is modeled in terms of weak supervision, while no prior knowledge is used. Specifically, in order to extract the label from the text for each instance of a visual concept, we face two distinct problems. (i) The first is the absence of specific spatio-temporal correspondence between visual and textual elements. In particular, in the tasks mentioned above, the descriptions are never provided with spatial boundaries and the temporal ones are usually imprecise. (ii) The second major issue is the semantic ambiguity of each textual element. This means that, when it comes to inferring complex semantics from the video such as actions or emotions, the extraction of the label from the text is no longer a straightforward procedure. For example, various expressions could be used to describe the action labeled as "walking", such as "lurching" or "going for a stroll".

Most of the work so far has dealt only to an extent with the spatio-temporal ambiguity, while the semantic one was totally ignored [7, 16, 23]. In this work, we introduce two novel weakly supervised techniques extending the Multiple Instance Learning (MIL)[1] discriminative clustering framework of [7]. The first one accounts for the temporal ambiguity variations, which are modeled by Fuzzy Sets (Fuzzy Sets MIL - FSMIL), while the second models the semantic ambiguities by probability mass functions (p.m.f) over the label set (Probabilistic Labels MIL - PLMIL). To the best of our knowledge, this is the first time that both methods are formulated in the context of MIL. In addition, we propose a method of extracting labels in complex tasks using semantic similarity computations. We further improve the recognition, from the perspective of visual representations using features learned from pre-trained deep architectures. The combination of all the above ideas leads to superior performance compared to previous work.

Finally, we focus on the recognition of faces and actions and the evaluation is performed on the COGNIMUSE database [49]. It is important to mention that our methods can be applied to other categories of concepts as long as they can be explicitly described in both modalities (video & text).

[1] In this paper, the term MIL does not concern only binary classification problems with positive and negative bags, as in its original definition [15], but also the multi-class case.

## 2. Related Work

During the last few years there have been various approaches of understanding videos or images using natural language. Specifically, many have approached the problem as machine translation, such as in [17], where image regions are matched to words of an accompanying caption and in [31, 38], where representations that translate video to sentences and vice-versa are learned. Others have tackled it using video-to-text alignment algorithms [8, 40].

Several works have considered text as means of supervision. In the problem of naming faces, Berg *et al*. [5, 6] use Linear Discriminant Analysis (LDA) followed by a modified k-means, to classify faces in newspaper images, while the labels are obtained from captions. In [7, 11, 12, 19, 32, 34, 42] the authors tackle a similar problem classifying faces in TV series or movies using the names of the speakers provided by the corresponding scripts. The proposed methods are based either on semi-supervised alike techniques using exemplar tracks [19, 42], ambiguous labeling [11, 12] or MIL [7, 32, 34].

The problem of automatically annotating actions in videos has recently drawn the attention of several researchers, because of the need to create diverse and realistic datasets of human activities. For this purpose, Laptev *et al*. used movie scripts to collect and learn realistic actions [23]. Later on, this work has been improved by incorporating information from the context, leading to the creation of the Hollywood2 dataset [27], and by a more accurate temporal localization using MIL [16]. In these, a Bag-of-Words text classifier is trained with annotated sentences in order to locate specific actions in the scripts. On the contrary, our work is based only on semantic similarity eliminating the cost of annotation. In Bojanowski *et al*. [7], MIL is also used to jointly learn names and actions, while in Miech *et al*. [28], the algorithm is improved allowing large-scale optimization via a variant of the Block-Coordinate Frank-Wolfe algorithm. In [28], a supervised approach is once more followed for the extraction of labels from the text, contrary to [7], where SEMAFOR [13] is used, a semantic role labeling parser, searching for two action frames. This unsupervised method, despite its promising results, cannot be easily generalized to custom actions. Similarly, in [41], the authors propose methods for learning multiple concepts jointly, introducing an extension of the Indian Buffer Process that is constrained by the information provided by the text. All the above end up in considering only the most certain labels that the text provides, ignoring possible paraphrases or synonyms. This allows an automatic collection of data with limited noise, but in general it leads to understanding

a small proportion of each individual video.

In order to learn from partially labeled data, there has been an extensive study on weakly supervised classification [21]. Learning with probabilistic labels has been examined in [22] under a probabilistic framework. Cour *et al.* [11] formulated a sub-category of this method, where all possible labels are distributed uniformly (candidate labels) and the classification is performed by minimizing a convex loss function. Both papers concern a single instance setting, namely a p.m.f over the label set is assigned to individual instances. On the contrary, we assign a p.m.f to bags-of instances, generalizing previous formulations. MIL has been largely studied in the machine learning community starting from Diettrich *et al.* [15], where drug activity was predicted. Except for the efforts on naming faces mentioned before, MIL has been used in detecting objects [48] and classifying scenes [26] in images, where annotation lacks specific spatial localization. While the definition of MIL is sufficient for most of its applications, it is important sometimes to make discriminations between instances in each bag. In order to model this case, we redefine MIL using Fuzzy Sets.

# 3. Multimodal Learning of Concepts

Given a video and a descriptive text that are crudely aligned [19], namely each portion of the text is attributed with temporal boundaries, we aim to localize and identify all the existing instances of a chosen visual concept, such as faces, actions or scenes. The adversities of such a task are clearly illustrated in Figure 1, where the concept examined is that of human actions. Our approach breaks down the problem into three subproblems. (a) First of all, the exact position in space and time of each visual concept is unknown, thus it needs to be detected automatically. (b) Secondly, concepts are usually expressed in the text in a different way than their original definition. For instance, as shown in Figure 1, the action "standing up" is mentioned by the phrase "gets up", while the action "answering phone" is mentioned by the phrase "opens his cell phone". In order to tackle this problem, we need to detect the part of the text that implies a concept and then mine the label information. (c) Finally, following the alignment procedure, the text is divided into segments that describe specific time intervals of the video. Each one of them might mention more than one instances of a visual concept. Thus, we need to apply a learning procedure that matches the mined labels with the detected concepts. Note here that sometimes a concept described in the text might not appear in the video or vice-versa. As a result, we need to design an algorithm that learns the visual concepts globally without restricting each one of them to the labels mentioned in its corresponding time interval.

Solving (a) and (b) requires task dependent systems, which are both described in section 4. The outputs of these

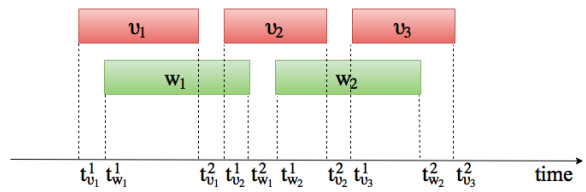



Figure 2: Illustration of the two modalities as parallel data streams.

systems are perceived as visual and linguistic objects ($v$ and $w$, respectively) with their temporal boundaries determined. Following the computation of these, we address (c) and we formulate the learning algorithms.

## 3.1. Problem Statement

We assume a dual modality scheme, where both modalities carry the same semantics. This can be modeled with two data streams flowing in parallel as time evolves (Figure 2). The first data stream consists of the unidentified visual objects that we want to recognize. We denote as $\mathcal{V}$ the set of $V$ visual objects. The second modality consists of the linguistic objects that carry in some way the information for the identification of each $v \in \mathcal{V}$, namely they describe the $v$. We denote as $\mathcal{W}$ the set of $W$ linguistic objects (*i.e.* words or sentences).

We assume that each $v$ is represented in a feature space and its representation is a vector $\boldsymbol{x}_v$. We define a matrix $\boldsymbol{X} \in \mathbb{R}^{V \times D}$ containing all the visual features. The time interval of each $v$ is denoted as $T_v = [t_v^1, t_v^2]$.

Let $\mathcal{Y}$ be the label set of $Y$ discrete labels. Each $w \in \mathcal{W}$ is mapped to a label $y_w$ through a mapping $\psi : \mathcal{W} \rightarrow \mathcal{Y}$. This can be either deterministic, matching each $w$ to a sole label $y$, or probabilistic, matching each $w$ to a p.m.f over the label set (see section 3.3.2). The time interval of each $w$ is denoted as $T_w = [t_w^1, t_w^2]$.

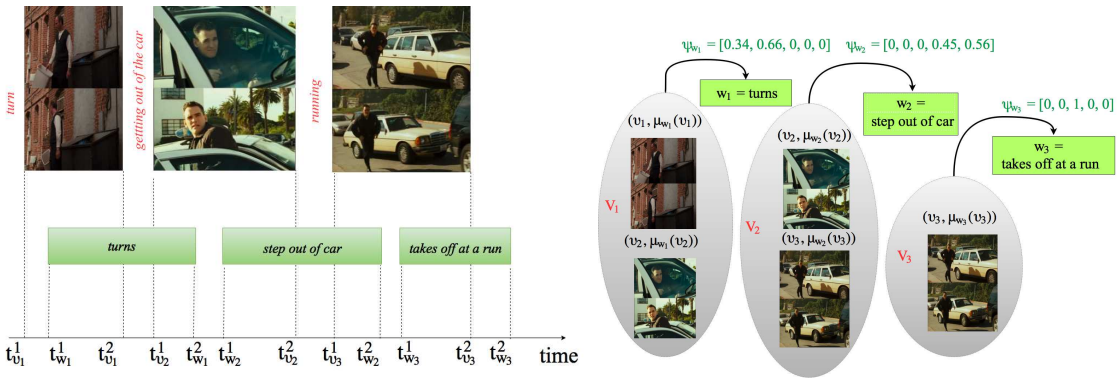
Our goal is to assign a specific label to each $v$, drawn from $\mathcal{Y}$. We denote the indicator matrix $\boldsymbol{Z} \in \{0,1\}^{V \times Y}$, which means that $z_{vy} = 1$ iff the label assigned to $v$ equals $y$. We want to infer $\boldsymbol{Z}$ given the visual feature matrix $\boldsymbol{X}$, the mapping $\psi$ and the temporal intervals $T_v$, $T_w$.

## 3.2. Clustering Model

Our model is based on DIFFRAC [4], a discriminative clustering method. In particular, Bach and Harchaoui, in order to assign labels to unsupervised data, form a ridge regression loss function using a linear classifier $f(\boldsymbol{x}) = \boldsymbol{x}\boldsymbol{\omega} + \boldsymbol{b}$, where $\boldsymbol{\omega} \in \mathbb{R}^{D \times Y}$ and $\boldsymbol{b} \in \mathbb{R}^{1 \times Y}$, which is optimized by the following:

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

where $\lambda$ stands for the regularization parameter. Eq. (1) can be solved analytically w.r.t. the classifier leading to a




(a) (b)

Figure 3: **(a)**: Example of the two data streams formed by linguistic and visual objects concerning the concept of human actions. Next to each visual object we demonstrate its ground truth. The formation of the streams is carried out by solving subproblems (a) and (b). **(b)**: The construction of bags under the MIL setting with Fuzzy Sets and Probabilistic Labels. The label set here is [*"walk", "turn", "running", "driving car", "getting out the car"*].

new objective function that needs to be minimized only w.r.t. the assignment matrix $\boldsymbol{Z}$: $\min_{\boldsymbol{Z}} \frac{1}{2}(\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda))$, where $\boldsymbol{A}(\boldsymbol{X}, \lambda)$ is a matrix that depends on the parameter $\lambda$ and the Gram matrix $\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}$, which can be replaced by any kernel (see [4]). Relaxing the matrix $\boldsymbol{Z} \in \{0,1\}^{V \times Y}$ to $\boldsymbol{Z} \in [0,1]^{V \times Y}$, the objective becomes a convex quadratic function constrained by the following:

$$\forall v \in \mathcal{V},\ \forall y \in \mathcal{Y},\ z_{vy} \geq 0, \tag{2}$$

$$\boldsymbol{Z} \cdot \boldsymbol{1}_Y = \boldsymbol{1}_V, \tag{3}$$

where $z_{vy}$ denotes the $(v, y)$ element of matrix $\boldsymbol{Z}$.

### 3.3. Weakly Supervised Methods

In order to incorporate in the model the weak supervision that the complementary modality provides, we have to resolve two kinds of ambiguities:

• Which visual object $v$ is described by each linguistic object $w$?

• Which label $y_w$ is implied from each $w$?

#### 3.3.1 Fuzzy Sets Multiple Instance Learning - FSMIL

In an attempt to address the first question, similar to [7], we assume that each $w$ should describe at least one of the $v$ that temporally overlaps with it. This leads to a multi-class MIL approach, where for each $w$ a bag-of-instances is created containing all the overlapping $v$:

$$\mathcal{V}_w = \{v \,|\, v \in \mathcal{V}\,,\, T_w \cap T_v \neq 0\}. \tag{4}$$

We extend this framework in order to discriminate between visual objects with different temporal overlaps. In fact, the longer the overlap, the more likely it is for a visual object $v$ to be described by the corresponding linguistic $w$. For example, during a dialogue, in the video stream the camera usually focuses on the current speaker longer than the silent person, while the document mentions the first. Thus, we need to encode this observation on our MIL sets. This is done by defining a novel type of MIL sets using fuzzy logic (see Figure 3). Each member of the set is accompanied by a value that demonstrates its membership grade :

$$\mathcal{V}_w = \{(v, \mu_w(v)) \,|\, v \in \mathcal{V}\,,\, \mu_w(v) = g(\frac{|T_w \cap T_v|}{|T_v|})\}, \tag{5}$$

where $g$ is an increasing membership function with $g(0) = 0,\ g(1) = 1$. In addition, we note that, in order to compensate for the crude alignment mistakes, we can add a hyper-parameter $\epsilon$ that adjusts the linguistic object time interval as follows: $\acute{T_w} = [t_w^1 - \epsilon r, t_w^2 + \epsilon r]$, where $r = |T_w| \,/\, |\overline{T_w}|$ and $|\overline{T_w}|$ is the average value of $|T_w|$, over all $w$.

#### 3.3.2 Probabilistic Labels Multiple Instance Learning - PLMIL

As mentioned before, the labels extracted from the complementary modality involve a level of uncertainty. This happens due to the fact that the extraction procedure is a classification problem on its own. Solving this problem is equivalent to inferring the mapping $\psi$. Obtaining the label that the classifier predicts for each linguistic object $w$, renders the mapping deterministic, while obtaining the posterior probabilities that the classifier gives, renders it probabilistic.

In this work, we use a probabilistic mapping using the posterior probabilities $\psi_w(y) = \mathbb{P}[y_w = y|w]$. In order to match them with the visual objects $v$, we perceive them as Probabilistic Labels (PLs). As mentioned in [22], matching a PL to an instance that needs to be classified, accounts for an initial estimation of its true label. In our problem, we generalize the definition of [22], matching PLs to bags-of-instances, meaning that at least one instance of the set

should be described by this measure of initial confidence. In this case, the model's input data is formed as follows: $D = \bigcup_{w \in \mathcal{W}} (\mathcal{V}_w, \psi_w(\cdot))$.

We address the classification problem of text segments in an unsupervised manner. Specifically, we calculate the semantic similarity $s_{wy}$ of each $w$ with the linguistic representation of each label $y$ using the algorithm of [20]. We also apply a threshold $\theta$ to each similarity value in order to eliminate the noisy $w$ that do not imply any of the labels. Thus, for each $w$ we obtain a similarity vector $\boldsymbol{s}_w$, which is then normalized to constitute a p.m.f : $\psi_w(y) = s_{wy} \,/\, \sum_{\ell \in \mathcal{Y}} s_{w\ell}$.

#### 3.3.3 Integration of the Weak Supervision in the Clustering Model

In the MIL case each bag $\mathcal{V}_w$ is matched to a single label $y$ and is represented by the following constraint:

$$\forall w \in \mathcal{W} \,,\ \sum_{v \in \mathcal{V}_w} z_{vy} \geq 1. \tag{6}$$

For the purpose of accounting for noise, slack variables are used to reformulate both the objective function and the constraints:

$$\min_{\boldsymbol{Z}, \boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda)) + \kappa \sum_{w \in \mathcal{W}} \xi_w^2, \tag{7}$$

$$\text{s.t} \quad \forall w \in \mathcal{W} \,,\ \sum_{v \in \mathcal{V}_w} z_{vy} \geq 1 - \xi_w. \tag{8}$$

In our FSMIL, we intend to add different weights to the elements of each bag depending on the membership grade:

$$\forall w \in \mathcal{W} \,,\ \sum_{v \in \mathcal{V}_w} z_{vy} \mu_w(v) \geq 1 - \xi_w. \tag{9}$$

For the PLMIL case, let $\mathcal{Y}_w \subseteq \mathcal{Y}$ be the set of the labels for which the p.m.f $\psi_w(y)$ is non-zero. For each label $y \in \mathcal{Y}_w$ we construct a constraint formed as in (9), *i.e.*:

$$\forall w \in \mathcal{W} \,,\ \forall y \in \mathcal{Y}_w \,,\ \sum_{v \in \mathcal{V}_w} z_{vy} \mu_w(v) \geq 1 - \xi_{wy}. \tag{10}$$

The discrimination between the various labels of $\mathcal{Y}_w$ is carried out by the slack variables. In particular, we rewrite the objective function as follows:

$$\min_{\boldsymbol{Z}, \boldsymbol{\xi}} Tr(\boldsymbol{Z}\boldsymbol{Z}^T \boldsymbol{A}(\boldsymbol{X}, \lambda)) + \kappa \sum_{w \in \mathcal{W}} \sum_{y \in \mathcal{Y}_w} f(\psi_w(y)) \xi_{wy}^2, \tag{11}$$

where $f(\cdot)$ is an increasing function that transforms probabilities to slack variable weights. In this way, we manage to relax the constraints inversely proportional to the probability of the corresponding label. As a result, a constraint is harder to be violated as long as the probability is high.

**Rounding:** Similarly to [7] we choose a simple rounding procedure for $\boldsymbol{Z}$ that accounts for taking the maximum values along its rows and replacing it with 1. The rest of the values are replaced with 0.

# 4. Experiments

## 4.1. Dataset

The COGNIMUSE database [1] is a video-oriented database multimodally annotated with audio-visual events, saliency, cross-media relations and emotion [49]. It is a generic database that can be used for event detection and summarization, as well as audio-visual concept recognition. Other existing databases such as the MPII-MD [37], the M-VAD [43], the MSVD [10], the MSR-VTT [46], the VTW [47], the TACoS [35, 39], the TACoS Multi-Level [36] and the YouCook [14] are not annotated in terms of specific visual concepts, but in terms of sentence descriptions. Moreover, the datasets used in [11, 19, 34, 42] are only annotated with human faces. Finally, the Hollywood2 [27] and the Casablanca [7] datasets were not sufficient for the action recognition task, due to the fact that only automatically collected labels from the text are provided rather than the text itself. Regarding Casablanca, it was not possible to apply our FMSIL method for the face recognition task either, as we observed that, in this movie, each face track and each sentence have either zero overlap ($T_v \cap T_w = \emptyset$) or the face track's temporal interval is a subset of that of the sentence (hence $T_v \cap T_w = T_v$), which always yields a zero or unitary membership grade for the FSMIL case. On the contrary, the COGNIMUSE database consists of long videos that are continuously annotated with action labels and are accompanied by texts in a raw format. In addition, we manually annotated the detected face tracks in order to evaluate the face recognition task. All the above, render COGNIMUSE more relevant and useful for the tasks that we are dealing with. In this work, we used 5 out of the 7 annotated 30-minutes movie clips, which are: A Beautiful Mind (BMI), Crash (CRA), The Departed (DEP), Gladiator (GLA) and Lord of the Rings - the Return of the King (LOR).

## 4.2. Implementation

**Detection and Feature Extraction**: We spatio-temporally detect and track faces similarly to [7], where face tracks are represented by SIFT descriptors and the kernels are computed separately for each facial feature taking into account whether a face is frontal or profile. Contrary to this, we use deep features extracted by the last fully connected layer of the VGG-face pre-trained CNN [33], while a single kernel is computed on each pair of face tracks regardless to the faces' poses. Similarly to [7, 19, 42] the kernel applied is a min-min RBF. For the problem of action recognition, we use the temporal boundaries provided by the dataset. We represent them through the C3D pre-trained CNN, following the methodology stated in [44] .

**Label Mining from Text**: Prior to applying the label extraction algorithms, we perform a crude alignment between the script and the subtitles through a widely used DTW al-

| 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** [7] | 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** |
| **[28]+VGG: fg** | 0.788 | 0.898 | 0.843 | 0.666 | 0.479 | 0.577 | 0.574 | 0.682 |
| **[28]+VGG+FSMIL: fg** | **0.810** | **0.913** | **0.862** | **0.696** | **0.505** | **0.651** | **0.617** | **0.715** |
| **[28]+VGG: bg** | **0.185** | **0.189** | **0.187** | **0.304** | 0.047 | **0.052** | 0.134 | 0.155 |
| **[28]+VGG+FSMIL: bg** | 0.184 | **0.189** | **0.187** | 0.269 | **0.278** | 0.038 | **0.195** | **0.192** |

Table 1: The Average Precision (AP) scores of each movie in the Development and the Test Set for the Face Recognition Task. The Mean Average Precision (MAP) calculated in the two sets separately and for the database as a whole is also shown.

gorithm [19]. The label set for the face recognition task is defined using the cast list of each movie (this information was downloaded from the Website TMDB [2]). The character labels are then extracted using regular expression matching, where the query expressions are the names included in the cast list. We define the label set for the action recognition task using a subset of the total classes of the COGNIMUSE database. We locate the linguistic objects $w$ by composing short sentences constituted by each sentence's verb as well as words that are linked to the verb through specific dependencies, such as the direct object and adverbs. We use the toolbox CoreNLP [25] in order to perform the document's dependency parsing. Finally, we calculate the semantic similarities on every label - short sentence pair applying an off-the-shelf sentence similarity algorithm [20]. This comprises a hybrid approach between Latent Semantic Analysis (LSA) and knowledge from WordNet [29]. The similarities that do not exceed a specific threshold $\theta$, experimentally set to 0.4, are discarded.

## 4.3. Learning Experiments

In the following experiments we evaluate our methods on the tasks of (i) face and (ii) action recognition. For the FSMIL setting, after extensive experimentation with a variety of $\Gamma$ and $S$-shaped membership functions (*e.g.* sigmoid, linear, piecewise quadratic, cubic), we selected a specific $\Gamma$-shaped function, using as criteria the performance and sensitivity to hyperparameters, that is defined as follows:

$$g(x) = \begin{cases} 0 & x \leq a \\ \frac{k(x-a)^2}{1+k(x-a)^2} & a \leq x \leq 1 \end{cases} \tag{12}$$

where $a$ is the membership threshold and $k$ is a parameter that controls how abrupt the increase above the threshold will be. We need to assign $k$ a large value (above 1000) in order to have $g(1) = 1$. For those values there are no significant changes in the results. We tune the hyperparameters $a$ and $\epsilon$ on the Development set independently for the two tasks, yielding $a = 0.2$, $\epsilon = 0$ for task (i) and $a = 0.1$, $\epsilon = 100$ for task (ii). For the PLMIL setting, we observed that, in the COGNIMUSE dataset, the best results were obtained when weights were close to 1 (candidate labels), thus the mapping function $f$ is again given by (12), setting $a$ to 0 and $k$ to a large value as previously stated. The hyperparameter sensitivity was assessed on the Test set and we saw that the deviations from the reported results were small, thus the sensitivity is assumed to be low. Moreover, the chosen values demonstrate almost optimal performance on the Test set as well, thus there was no overfitting on the Development set. The source code of the entire system and experiments (including results on hyperparameter sensitivity), as well as precomputed features can be found at [3].

### 4.3.1 Face Recognition

We evaluate each method's performance using the Average Precision (AP) previously used in [7, 11]. We compare our model (VGG+FSMIL) to the methodology of Bojanowski *et al.* [7] - that has outperformed other weakly supervised methods, such as [11] and [42] - as well as with other baselines described next:

1. Text+MIL: We solve the problem by minimizing only the factor related to the slack variables. This method, converges to an optimum that satisfies the constraints posed by the text, without taking into account the visual features. The constraints are formed using the simple MIL setting.
2. SIFT+MIL [7]: The algorithm of Bojanowski *et al.* that uses SIFT descriptors as feature vectors and simple MIL setting, without taking into account the temporal overlaps, namely the bags are constructed as noted by (4).
3. SIFT+FSMIL: Our proposed learning method implemented with SIFT descriptors.
4. VGG+MIL: The algorithm of Bojanowski *et al.* implemented with VGG-face descriptors.
5. **VGG+FSMIL (Ours)**: The proposed learning method implemented with VGG-face descriptors.

The comparative results for each movie are shown in Table 1. As it can be clearly seen, our method demonstrates superior performance than [7] concerning every case.

First of all, the inferior performance of the Text+MIL method shows the inefficiency of using only textual information in tackling the problem. The higher accuracy accomplished by the methods implemented with VGG proves the benefits of deep learning over hand-crafted features as means of representing faces. Moreover, incorporating the information given by the overlaps of visual and linguistic objects, we improve the accuracy regardless of the nature of the representation. In particular, due to the fact that our method reduces the ambiguity in each bag-of-instances, we outperform the baseline even without the use of deep features. As expected, the combination of the above (VGG and FSMIL) shows the highest accuracy. This can be easily explained as each one of the methods improves different aspects of the learning procedure.

In the comparisons above, we discard background (bg) characters and optimize on foreground (fg). Additionally, in Table 1, we compare our method to [28], that has achieved state-of-the-art results by using a bg class constraint. In particular, tracks that have no assigned weak label are collected in a bag and a percentage of them, determined by a hyper-parameter $\alpha_2$ (we set it to 0.6), is enforced by a constraint to be classified as bg. We observe that when we incorporate FSMIL, our method always outperforms [28] on fg characters and provides either similar, or better results on bg.

#### 4.3.2 Action Recognition

Regarding the task of action recognition, several experiments were carried out for each movie, while changing the cardinality of the label set. In particular, the performance is evaluated using the 2, 4, 6, 8 and 10 most frequent action classes. The evaluation is performed using the Mean AP metric, which stands for averaging the APs over each movie set. The results are demonstrated both for the Development and the Test set in Table 2. We also illustrate the performances of the methods on the whole dataset with the *per sample accuracy* vs *proportion of total instances* curves of Figure 4.

We choose as baseline the aforementioned Text+MIL, as well as a similar methodology to Bojanowski *et al*. [7]. In this experiment, we focus on the different ways of learning from the text, rather than the visual features, thus in all cases we use the C3D descriptor for the representation of actions. The methods compared are:

1. Text+MIL: Same as the one described in section 4.3.1. The action labels are extracted by locating the sentences that are **semantically identical** to one of the labels of the set $\mathcal{Y}$ (similarity = 1).
2. MIL ([7] modified): The learning algorithm of Bojanowski *et al*. mentioned in 4.3.1. We replace the dense trajectories descriptors with C3D. Again, we use only the sentences that are **semantically identical** to some label.
3. Sim+MIL: The same learning algorithm, but labels are extracted from sentences that are **semantically similar** to one of the labels of the set $\mathcal{Y}$ ($\theta \leq$ similarity $\leq 1$). Each sentence is assigned a single label, the one with the maximum similarity.
4. Sim+PLMIL: Our PLMIL method. We assign a probabilistic label to each sentence.
5. Sim+FSMIL: Our FSMIL method. We construct the bags-of-instances as fuzzy sets.
6. **Sim+FSMIL+PLMIL (Ours)**: The combination of our contributions using semantically similar sentences, probabilistic labels and fuzzy bags-of-instances.

| **Number of Classes** | **2** | **4** | **6** | **8** | **10** |
|---|---|---|---|---|---|
| **Set** | **Development** | | | | |
| **Text+MIL** | 0.566 | 0.315 | 0.253 | 0.083 | 0.089 |
| **MIL(modified) [7]** | 0.638 | 0.420 | 0.283 | 0.266 | 0.194 |
| **Sim+MIL** | 0.837 | 0.299 | 0.243 | 0.339 | 0.202 |
| **Sim+PLMIL** | 0.837 | 0.304 | 0.308 | 0.348 | 0.229 |
| **Sim+FSMIL** | **0.945** | 0.614 | 0.435 | 0.383 | 0.317 |
| **Sim+FSMIL+PLMIL (Ours)** | **0.945** | **0.617** | **0.520** | **0.491** | **0.450** |
| **Set** | **Test Set** | | | | |
| **Text+MIL** | 0.306 | 0.199 | 0.188 | 0.134 | 0.147 |
| **MIL(modified) [7]** | 0.405 | 0.180 | 0.184 | 0.189 | 0.212 |
| **Sim+MIL** | 0.631 | **0.591** | 0.182 | 0.094 | 0.129 |
| **Sim+PLMIL** | 0.585 | 0.564 | 0.298 | 0.140 | 0.148 |
| **Sim+FSMIL** | **0.792** | 0.458 | 0.232 | 0.168 | 0.146 |
| **Sim+FSMIL+PLMIL (Ours)** | 0.692 | 0.521 | **0.299** | **0.249** | **0.270** |

Table 2: The Mean Average Precision (AP) scores over the **Development** and **Test** set for five independent experiments for the Action Recognition Task.

First note that the proposed combined model demonstrates superior performance over the Text + MIL baseline, confirming the importance of using visual information, as previously mentioned in 4.3.1. Higher performance is also reported over the baseline of [7] in every case, leading to an improvement of 20% – 30% in the Development set and 6% – 34% in the Test set. Moreover, Figure 4 shows that it outperforms all methods in the whole dataset, except for the case of two classes. Next, we examine each of our contributions independently.

The method of extracting labels through similarity measurements outperforms the baseline mainly when the number of classes is small (2-4), as shown in Table 2. In this case, the concepts implied by the labels, in terms of semantics, are rarely confused, hence most of the similarity measurements produce correct labels. However, as this number increases the Sim+MIL method does not prove very efficient on its own. A possible explanation is that the semantically identical labels of the baseline usually consist of a more clean set, while the confusion introduced to the model with semantically similar labels rises. As a result, despite the fact that a small amount of bags-of-instances are annotated, the baseline algorithm will still be able to make a few correct predictions with large confidence. This is illustrated in Figure 4 (c) and (d), where the most confident predictions

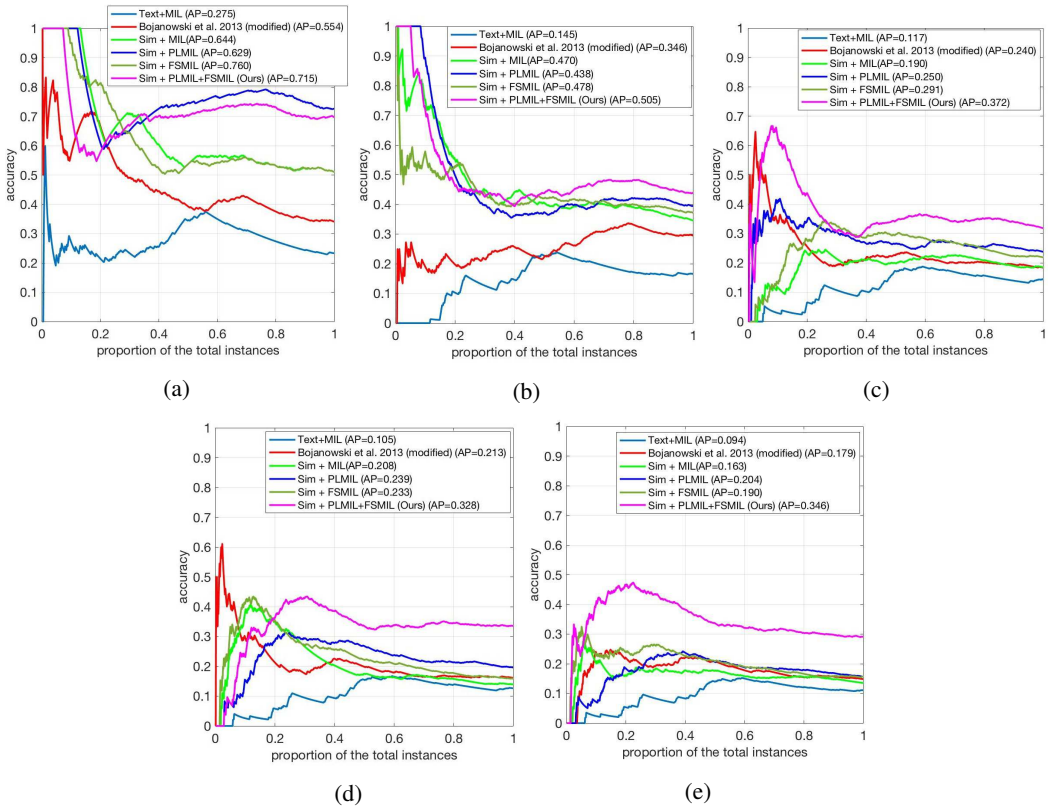



Figure 4: The curves show the per sample accuracy plotted against the proportion of total instances, concerning the whole dataset. Each figure corresponds to a different experiment concerning the number of classes. Note that as the number of classes increases, the model that combines our methods greatly outperforms each one of them individually as well as the baseline. Please see color version for better visibility.

of the baseline are accurate, contrary to those of Sim+MIL.

This confusion is compensated partially by either PLMIL or FSMIL. Regarding the first one, when the classes are few, a sentence is rarely similar to more than one concepts, hence the labels are mainly deterministic. However, modeling labels in a probabilistic way achieves better disambiguation of the sentences' meanings as the number of classes grows larger, which is proved by the fact that Sim+PLMIL outperforms Sim+MIL for 6-10 classes in both sets. As far as the FSMIL is concerned, this method is expected to perform better on its own for the reasons mentioned in section 4.3.1, regardless of the number of classes. Indeed, Sim+FSMIL outperforms Sim+MIL in most of the cases.

Interestingly, the combination of our contributions manages to outperform the baseline, even if none of them could do so independently. This can be explained by the fact that the algorithm leverages each one of them to resolve different kinds of ambiguities. Regarding the lower results in the Test set compared to the Development, we noticed that the scripts of the test movies are not sufficiently aligned to the videos, while a significant amount of actions occur in the background, consequently are not described in the text.

## 5. Conclusion

In this work we tackled the problem of automatically learning visual concepts by combining visual and textual information. We proposed two novel weakly supervised techniques that can be easily generalized to other Multimodal Learning tasks, that efficiently deal with temporal ambiguities (FSMIL), as well as semantic ones (PLMIL). Contrary to previous work, we acquire richer information from the text using semantic similarity. We evaluated our models on the COGNIMUSE dataset, containing densely annotated movies accompanied by their scripts. Our techniques provide significant improvement over a state-of-the-art weakly supervised method, in both face and action recognition tasks. Regarding our future work, we plan to extend our uni-directional model to a bi-directional, where information will flow from text to video and vice-versa, jointly learning visual and linguistic concepts. Finally, the generality of our formulation motivates us in exploring its potential in learning from other modalities such as the audio channel.

**Acknowledgements:** This work was partially supported by EU projects Babyrobot/687831 and i-Support/643666.

## References


[1] http://cognimuse.cs.ntua.gr/database. 5

[2] https://www.themoviedb.org/. 6

[3] https://github.com/gbouritsas/cvpr18_multimodal_weakly_supervised_learning. 6

[4] F. R. Bach and Z. Harchaoui. DIFFRAC: a discriminative and flexible framework for clustering. In *NIPS*, 2008. 3, 4

[5] T. L. Berg, A. C. Berg, J. Edwards, and D. A. Forsyth. Who's in the picture. In *NIPS*, 2005. 2

[6] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y.-W. Teh, E. Learned-Miller, and D. A. Forsyth. Names and faces in the news. In *CVPR*, 2004. 2

[7] P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, and J. Sivic. Finding actors and actions in movies. In *ICCV*, 2013. 2, 4, 5, 6, 7

[8] P. Bojanowski, R. Lajugie, E. Grave, F. Bach, I. Laptev, J. Ponce, and C. Schmid. Weakly-supervised alignment of video with text. In *ICCV*, 2015. 2

[9] H. Bredin, C. Barras, and C. Guinaudeau. Multimodal person discovery in broadcast tv at mediaeval 2016. In *MediaEval*, 2016. 2

[10] D. L. Chen and W. B. Dolan. Collecting highly parallel data for paraphrase evaluation. In *ACL*, 2011. 5

[11] T. Cour, B. Sapp, C. Jordan, and B. Taskar. Learning from ambiguously labeled images. In *CVPR*, 2009. 2, 3, 5, 6

[12] T. Cour, B. Sapp, and B. Taskar. Learning from partial labels. *JMLR*, 2011. 2

[13] D. Das, A. F. Martins, and N. A. Smith. An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *SEM*, 2012. 2

[14] P. Das, C. Xu, R. F. Doell, and b. J. J. Corso. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *CVPR*, 2013. 5

[15] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artif Intell*, 1997. 2, 3

[16] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009. 2

[17] P. Duygulu, K. Barnard, J. F. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV*, 2002. 2

[18] G. Evangelopoulos, A. Zlatintsi, A. Potamianos, P. Maragos, K. Rapantzikos, G. Skoumas, and Y. Avrithis. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE TMM*, 2013. 2

[19] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006. 2, 3, 5, 6

[20] L. Han, A. Kashyap, T. Finin, J. Mayfield, and J. Weese. Umbc ebiquity-core: Semantic textual similarity systems. In **SEM*, 2013. 5, 6

[21] J. Hernández-González, I. Inza, and J. A. Lozano. Weak supervision and other non-standard classification problems: a taxonomy. *Pattern Recogn Lett*, 2016. 3

[22] R. Jin and Z. Ghahramani. Learning with multiple labels. In *NIPS*, 2003. 3, 4

[23] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2

[24] S. Maji and R. Bajcsy. Fast unsupervised alignment of video and text for indexing/names and faces. In *Workshop on multimedia information retrieval on The many faces of multimedia semantics*, 2007. 2

[25] C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard, and D. McClosky. The Stanford CoreNLP natural language processing toolkit. In *ACL*, 2014. 6

[26] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *ICML*, 1998. 3

[27] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 5

[28] A. Miech, J.-B. Alayrac, P. Bojanowski, I. Laptev, and J. Sivic. Learning from video and text via large-scale discriminative clustering. In *ICCV*, 2017. 2, 6, 7

[29] G. A. Miller. Wordnet: a lexical database for english. *ACM*, 1995. 6

[30] T. S. Motwani and R. J. Mooney. Improving video activity recognition using object recognition and text mining. In *ECAI*, 2012. 2

[31] S. Naha and Y. Wang. Beyond verbs: Understanding actions in videos with text. In *ICPR*, 2016. 2

[32] O. M. Parkhi, E. Rahtu, and A. Zisserman. It's in the bag: Stronger supervision for automated face labelling. In *ICCV Workshop: Describing and Understanding Video & The Large Scale Movie Description Challenge*, 2015. 2

[33] O. M. Parkhi, A. Vedaldi, A. Zisserman, et al. Deep face recognition. In *BMVC*, 2015. 5

[34] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei. Linking people in videos with their names using coreference resolution. In *ECCV*, 2014. 2, 5

[35] M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal. Grounding action descriptions in videos. *TACL*, 2013. 5

[36] A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, and B. Schiele. Coherent multi-sentence video description with variable level of detail. In *GCPR*, 2014. 5

[37] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele. A dataset for movie description. In *CVPR*, 2015. 5

[38] A. Rohrbach, A. Torabi, M. Rohrbach, N. Tandon, C. Pal, H. Larochelle, A. Courville, and B. Schiele. Movie description. *IJCV*, 2017. 2

[39] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *ICCV*, 2013. 5

[40] P. Sankar, C. V. Jawahar, and A. Zisserman. Subtitle-free movie to script alignment. In *BMVC*, 2009. 2

[41] S. Shah, K. Kulkarni, A. Biswas, A. Gandhi, O. Deshmukh, and L. S. Davis. Weakly supervised learning of heterogeneous concepts in videos. In *ECCV*, 2016. 2

[42] J. Sivic, M. Everingham, and A. Zisserman. who are you?-learning person specific classifiers from video. In *CVPR*, 2009. 2, 5, 6

[43] A. Torabi, C. Pal, H. Larochelle, and A. Courville. Using descriptive video services to create a large data source for video annotation research. *arXiv:1503.01070*, 2015. 5

[44] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 5

[45] Z. Wang, K. Kuan, M. Ravaut, G. Manek, S. Song, F. Yuan, K. Seokhwan, N. Chen, L. F. D. Enriquez, L. A. Tuan, et al. Truly multi-modal youtube-8m video classification with video, audio, and text. *arXiv:1706.05461*, 2017. 2

[46] J. Xu, T. Mei, T. Yao, and Y. Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 5

[47] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun. Generation for user generated videos. In *ECVV*, 2016. 5

[48] C. Zhang, J. C. Platt, and P. A. Viola. Multiple instance boosting for object detection. In *NIPS*, 2006. 3

[49] A. Zlatintsi, P. Koutras, G. Evangelopoulos, N. Malandrakis, N. Efthymiou, K. Pastra, A. Potamianos, and P. Maragos. COGNIMUSE: A multimodal video database annotated with saliency, events, semantics and emotion with application to summarization. *EURASIP Journal on Image and Video Processing*, 2017(1):54, Aug 2017. 2, 5