Classification d'images télévisées

Farhad - Heybati
Youcef - Kacer
Martin - Provost

9 May 2016

Table des matières

Introduction

Ce document présente plusieurs méthodes d'extraction d'informations à partir d'images issues d'un débat télévisé. Cela afin d'en effectuer une classification binaire en « gros plan » ou « large plan ». Nous proposons dans un premier chapître de présenter les images exploitées, et les deux classes qui nous intéressent. Dans une seconde partie, nous presentons les méthodes d'extraction d'informations utilisées. Puis dans une troisième partie, les résultats obtenus.

Chapitre 1

Images exploitées

1.1 Corpus et labelisation

Nous allons exploités un total de 2351 images prises à partir de la vidéo d'un débat télévisé [Ess16] . Nous avons exploité le fichier de transcription au format .trs [Ess16] associé à cette vidéo, cela afin d'extraire les différents classes, et attribuer à chaque image sa classe. En effet, le fichier de transcription fournit un total de 9 classes :

M **:** La présentatrice est seule à l'écran

A **:** La première intervenante est seule à l'écran

B **:** La seconde intervenante est seule à l'écran

C **:** Le premier intervenant est seul à l'écran

D **:** Le second intervenant est seul à l'écran

ALL **:** Les 5 personnes sont à l'écran

$MULTI$ **:** Entre 2 et 4 personnes sont à l'écran

$INTRO$ **:** Reportage d'introduction à l'écran

$CREDITS$ **:** Générique d'emission à l'écran

Par ailleurs, le fichier de transcription donne à chaque intervalle de temps, ce qui est à l'écran parmi les classes citées plus-haut. Moyennant, une conversion du fichier au format xml, on peut obtenir le tableau suivant :

classe	**debut (s)**	**fin (s)**
$CREDITS$	0	11.36
$INTRO$	11.36	84.64
M	84.64	95.12
ALL	95.12	103.2
A	103.2	112.2
M	112.2	115.24
$\vdots$	$\vdots$	$\vdots$
M	2334.12	2340
ALL	2340	2340.76
$CREDITS$	2340.76	2349.72

TABLE 1.1 – Table de correspondance classes/intervalle de temps

Les 2351 images étant prises à une seconde d'intervalle tout le long de la vidéo, on peut automatiquement labélisé celles ci via la table de correspondance 1.1. Ci-après, nous présentons quelques images pour chacune des 9 classes :

1.2 Classification binaire

Par la suite, nous allons nous resteindre à seulement deux classes définies comme suit :

G **:** « Gros plan » (une seule personne est à l'ecran)
L **:** « Large plan » (au moins deux personnes sont à l'écran)

Les classes G et L peuvent s'exprimer en fonction des 9 classes comme suit :

G **:** $M \mid A \mid B \mid C \mid D$
L **:** $ALL \mid MULTI$

Nous avons donc deux classes d'images G et L dont voici plusieurs exemples :

Par la suite, nous allons expliciter différentes manières d'extraire de l'information afin de pouvoir discriminer les images de classe G , des images de classe L.

FIGURE 1.1 – image de classe M

FIGURE 1.2 – image de classe A

FIGURE 1.3 – image de classe B

FIGURE 1.4 – image de classe C

FIGURE 1.5 – image de classe D

FIGURE 1.6 – image de classe ALL

FIGURE 1.7 – image de classe $MULTI$

FIGURE 1.8 – image de classe $INTRO$

FIGURE 1.9 – image de classe *CREDITS*

FIGURE 1.10 – image de classe G (gros plan)

FIGURE 1.11 – image de classe L (large plan)

Chapitre 2

Extraction d'information

2.1 Extraction de contours

2.2 Extraction de teinte

2.3 Histogramme orienté du gradient

Cette méthode consiste à extraire l'information local de contours. Sa première utilisation a consisté en la détection de piétons [DT05]. Pour une image donné, le vecteur descripteur des *HOG* est la concaténation d'histogrammes de l'amplitude du gradient (en fonction de l'orientation) pris sur un découpage de l'image :

On peut espérer que ces descripteurs s'adaptent bien à notre problème. En effet, les epaules des intervenants à gauche et à droite de l'image sont des contours descriminants pour la classe G (gros plan). D'autre part, les gros plans ont fond uniforme, ce qui donnera beaucoup de gradient nul, et donc un descripteur *sparse*

2.4 Reseaux de neurones à convolution pré-entrainé

Cette méthode consiste à extraire les descripteurs produits par un réseau de neurones à convolution déjà entrainé. En effet, en récupérant la sortie de l'avant-dernière couche, on obtient la transformation finale haut-niveau de l'image, avant la couche de classification. Cette technique est connu sous le nom de *Transfer Learning* [YCBL14] et a fait ses preuves dans le cas de réseaux de neurones entrainé sur la base d'images *ImageNet* [DDS$^+$09]. Cette base de près de 10 millins d'images, contient une multitude de classes (animaux,ustensiles,humains,...) organisé hiérarchiquement, et on peut espérer que les gros plans (classe G) correspondent à l'une des ses classes,et de même pour les images plan large (classe L), de telle sorte qu'on puisse ensuite séparer les descripteurs via un classifieur supervisé.

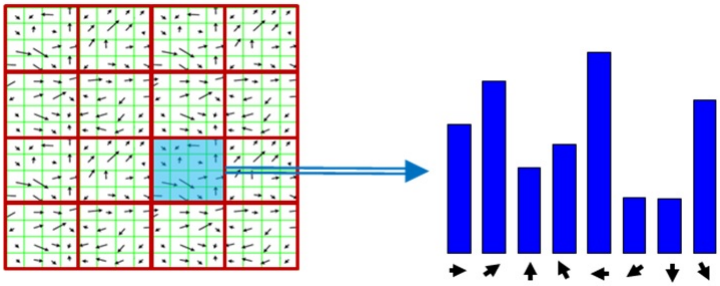

FIGURE 2.1 – construction d'un histogramme orienté du gradient [Lev13]

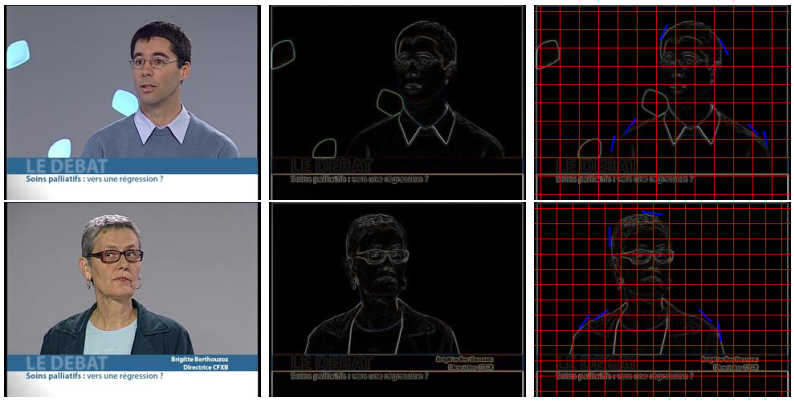

FIGURE 2.2 – histogramme orienté du gradient pour les gros plans (classe G)

Chapitre 3

Résultats

3.1 Extraction de contours

3.2 Extraction de teinte

3.3 Histogramme orienté du gradient

Nous avons extrait les descripteurs *HOG* pour chacune des images de classe G et L, en utilisant une fenêtre glissante de taille $32x32$ avec un overlap. Puis, nous leur avons appliquer différent classifieurs supervisés. Nous avons découpé le set de descripteurs en deux sous-ensembles représentant 70% du total pour l'entrainement, 30% pour le test. Les resultats de classification sont illustrés dans la table.

3.4 Reseaux de neurones à convolution pré-entrainé

Nous avons utilisé le code *Overfeat* [SEZ+13], récupéré par clonage du dépôt Github correspondant [SEZ+14], pré-entrainé sur la base *ImageNet* [DDS+09]. Nous avons extrait les descripteurs pour chacune des images de classe G et L pour leur appliquer différent classifieurs supervisés. Nous avons découpé le set de descripteurs en deux sous-ensembles s 70% du total pour l'entrainement, 30% pour le test. Les resultats de classification sont illustrés dans la table.

Liste des tableaux

Table des figures

Bibliographie

[DDS+09] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet : A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. *cvpr*, jun 2005.

[Ess16] Slim Essid. Resources, 2016. www.perso.telecom-paristech.fr/ essid/.

[Lev13] Gil Levi. A short introduction to descriptors, aug 2013. https ://gilscvblog.com.

[SEZ+13] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat : Integrated recognition, localization and detection using convolutional networks. *CoRR*, abs/1312.6229, 2013.

[SEZ+14] Pierre Sermanet, David Eigen, Xiang Zhang, Michaël Mathieu, Rob Fergus, and Yann LeCun. Overfeat code source, 2014. https ://github.com/sermanet/OverFeat.

[YCBL14] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks ? *CoRR*, abs/1411.1792, 2014.