Deep Learning en la Clasificación de Vídeos para Cine Digital

Umair A. Khan, Naveed Ejaz,

Miguel A. Martínez-del-Amor, Heiko Sparenberg

ERCIM

Fraunhofer

About me

- Miguel Ángel Martínez del Amor
  - Dpto. Ciencias de la Computación e Inteligencia Artificial
  - Universidad de Sevilla
  - *Web e info de contacto: www.cs.us.es/~mdelamor*

- Más información del entorno donde se realizó este trabajo:
  - Institutos Fraunhofer
  - Fraunhofer IIS
  - Grupo Cine Digital en Fraunhofer IIS
  - Becas posdoctorales ERCIM

Agenda

- Transfer Learning (conceptos básicos)
- Cine Digital
- Casos de estudio:
  - Tag extraction
  - Beat Event Detection

Agenda

- **Transfer Learning (conceptos básicos)**
- Cine Digital
- Casos de estudio:
  - Tag extraction
  - Beat Event Detection

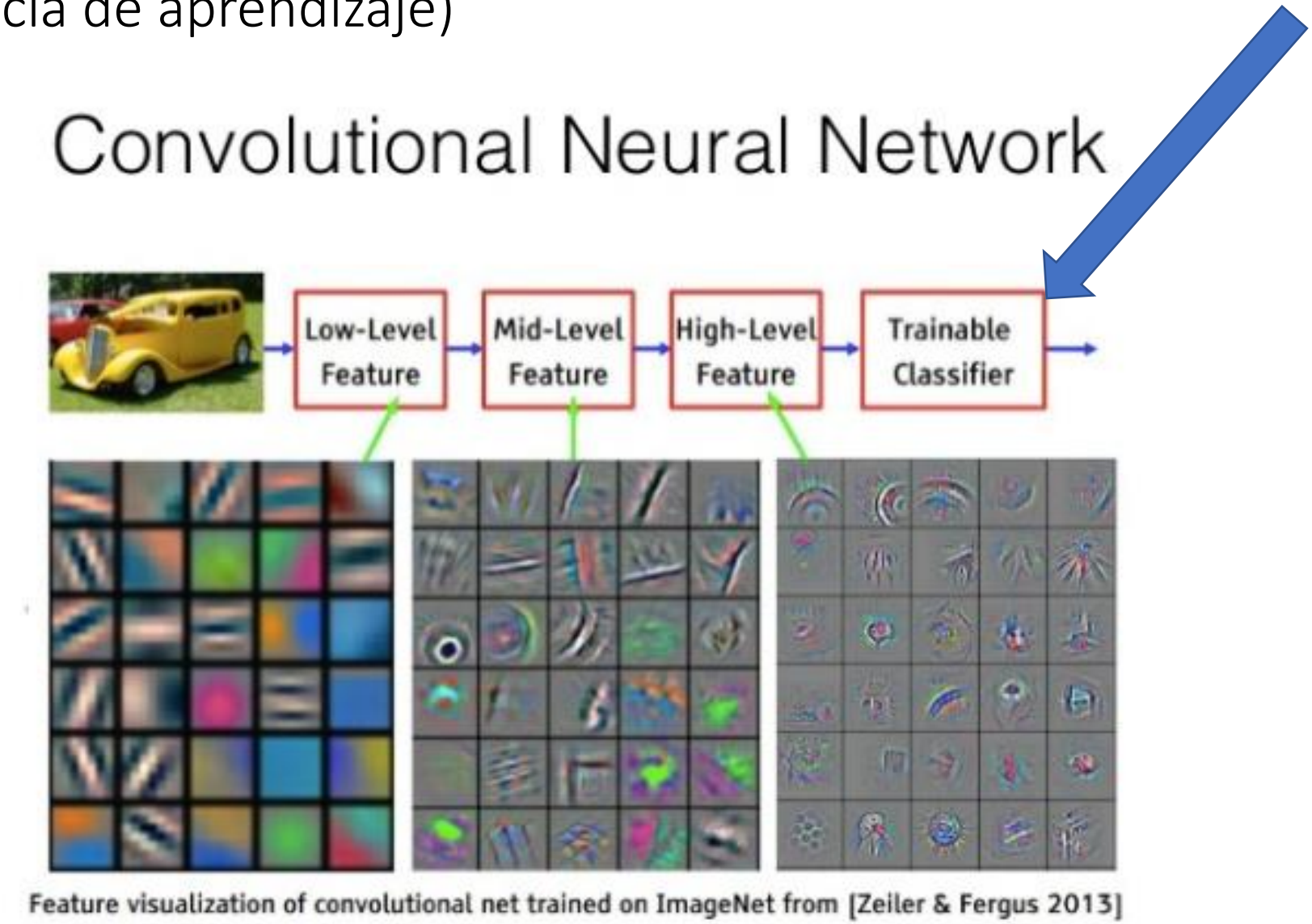
Transfer Learning
(transferencia de aprendizaje)

- Para entrenar una CNN desde cero se necesita:
  - **Much(ísim)os datos** (*p.ej. ImageNet: 1,2 millones de imágenes, 1000 categorías*)
  - Gran **capacidad computacional** (*p.ej. DGX v2 con 16 Tesla V100*).
  - **Tiempo** (*semanas a meses para entrenamiento*)

- En la realidad, pocos investigadores entrenan una CNN desde cero
  - Partir de una ConvNet pre-entrenada en un conjunto de datos muy grande

- Yosinski et al. How transferable are features in deep neural networks? 2014

Transfer Learning
(transferencia de aprendizaje)

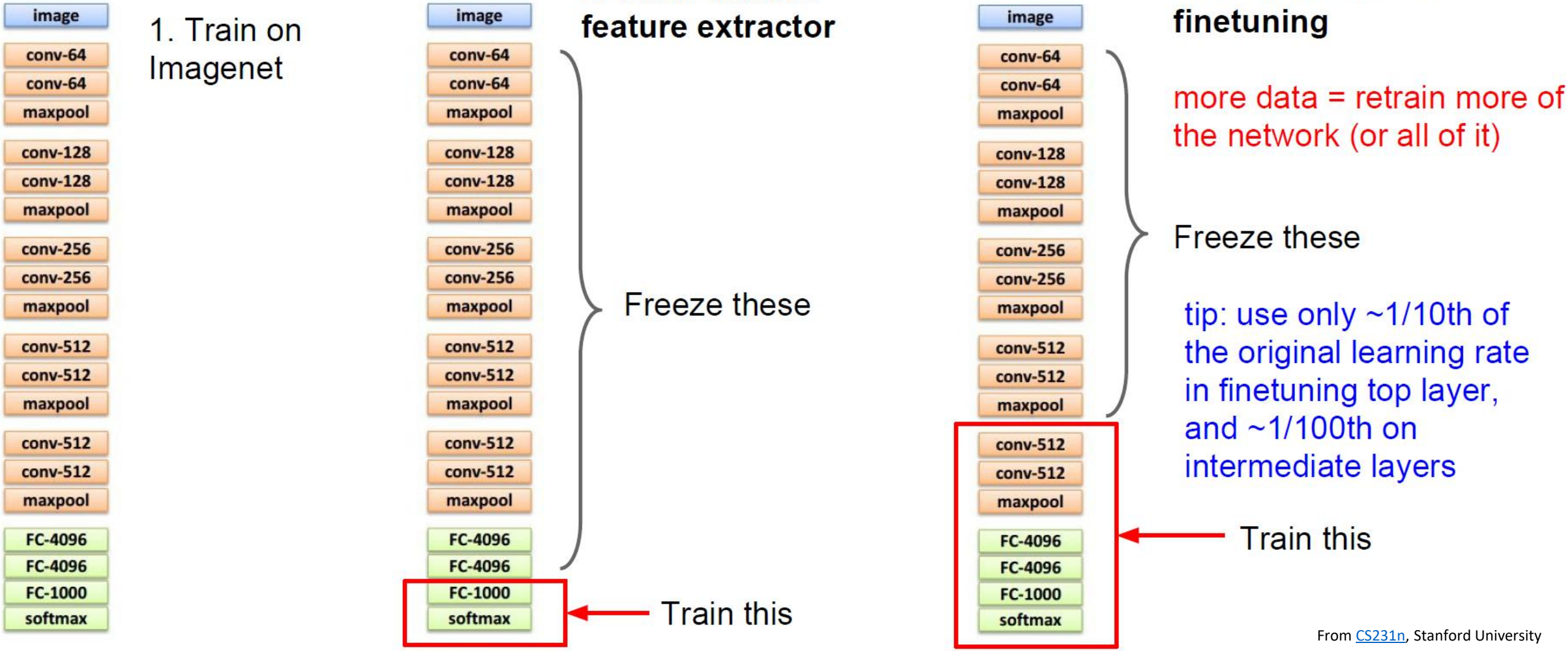


Transfer Learning
(transferencia de aprendizaje)

- Tres escenarios:
  - **Fixed feature extractor**: eliminar las últimas capas FC (Fully Connected) y el clasificador, y fijar el resto.
    - Re-entrenar el clasificador con el nuevo conjunto de datos.
  - **Fine-tuning**: fijar solo las primeras capas, aplicar *backpropagation* al resto.
    - Últimas capas suelen contener características más específicas a las categorías por las que fueron entrenadas.
  - **Pretrained models**: descargar un modelo (p.ej. *Model Zoo de Caffe*) para aplicar lo anterior.

http://cs231n.github.io/transfer-learning/

Transfer Learning



From CS231n, Stanford University

Transfer Learning
(transferencia de aprendizaje)

	very similar dataset	**very different dataset**
very little data	Use Linear Classifier on top layer	You're in trouble… Try linear classifier from different stages
quite a lot of data	Finetune a few layers	Finetune a larger number of layers

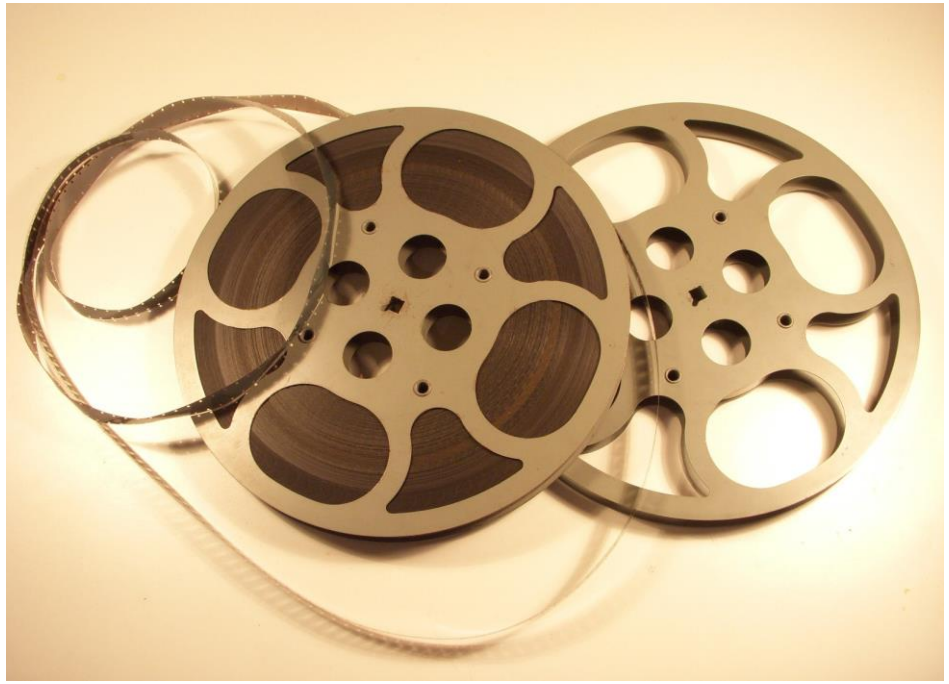
From CS231n, Stanford University

Agenda

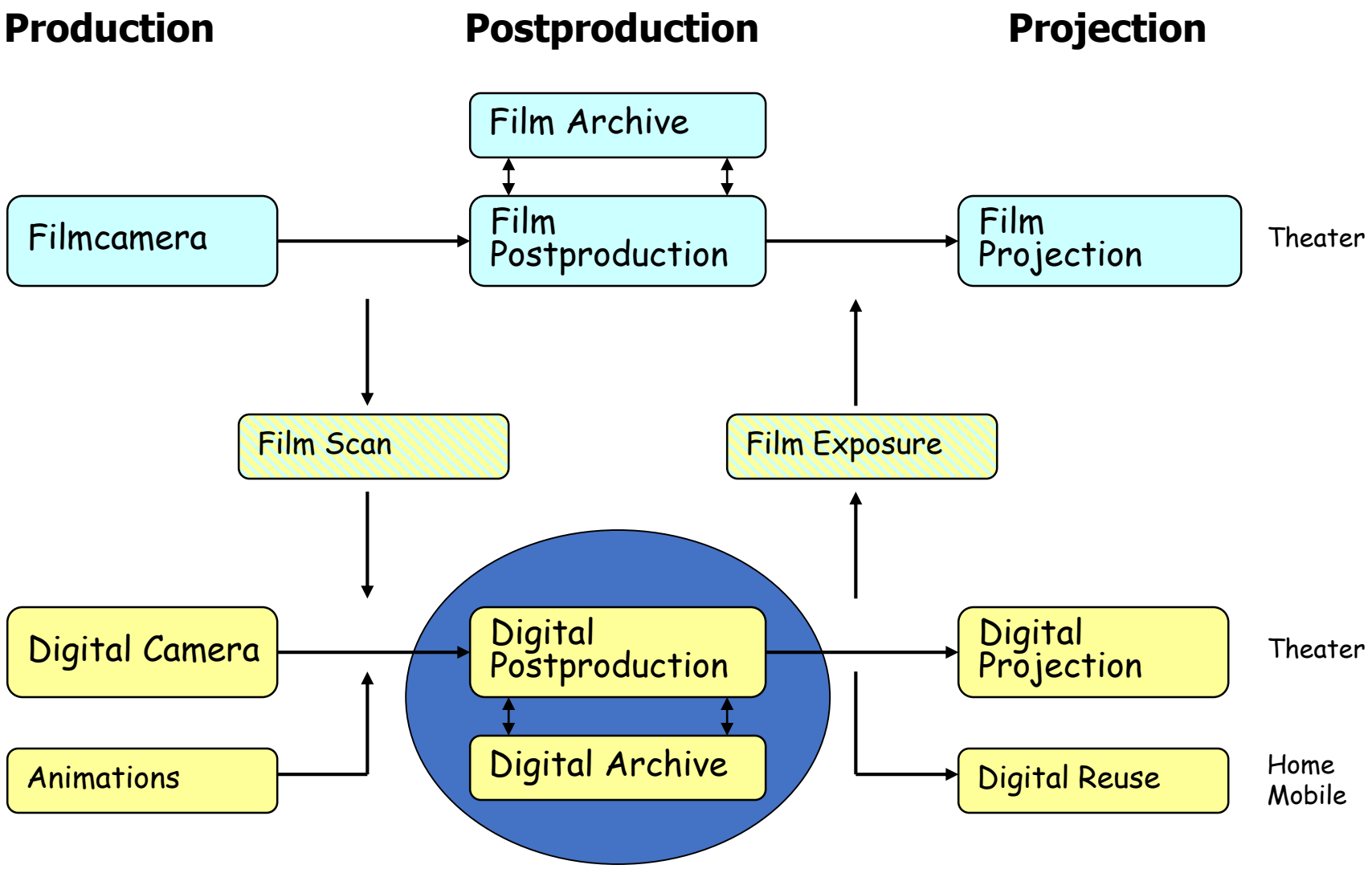
- Transfer Learning (conceptos básicos)
- **Cine Digital**
- Casos de estudio:
  - Tag extraction
  - Beat Event Detection

Cine Digital

- Digitalización completa de la proyección de contenido comercial (de la grabación al cine)

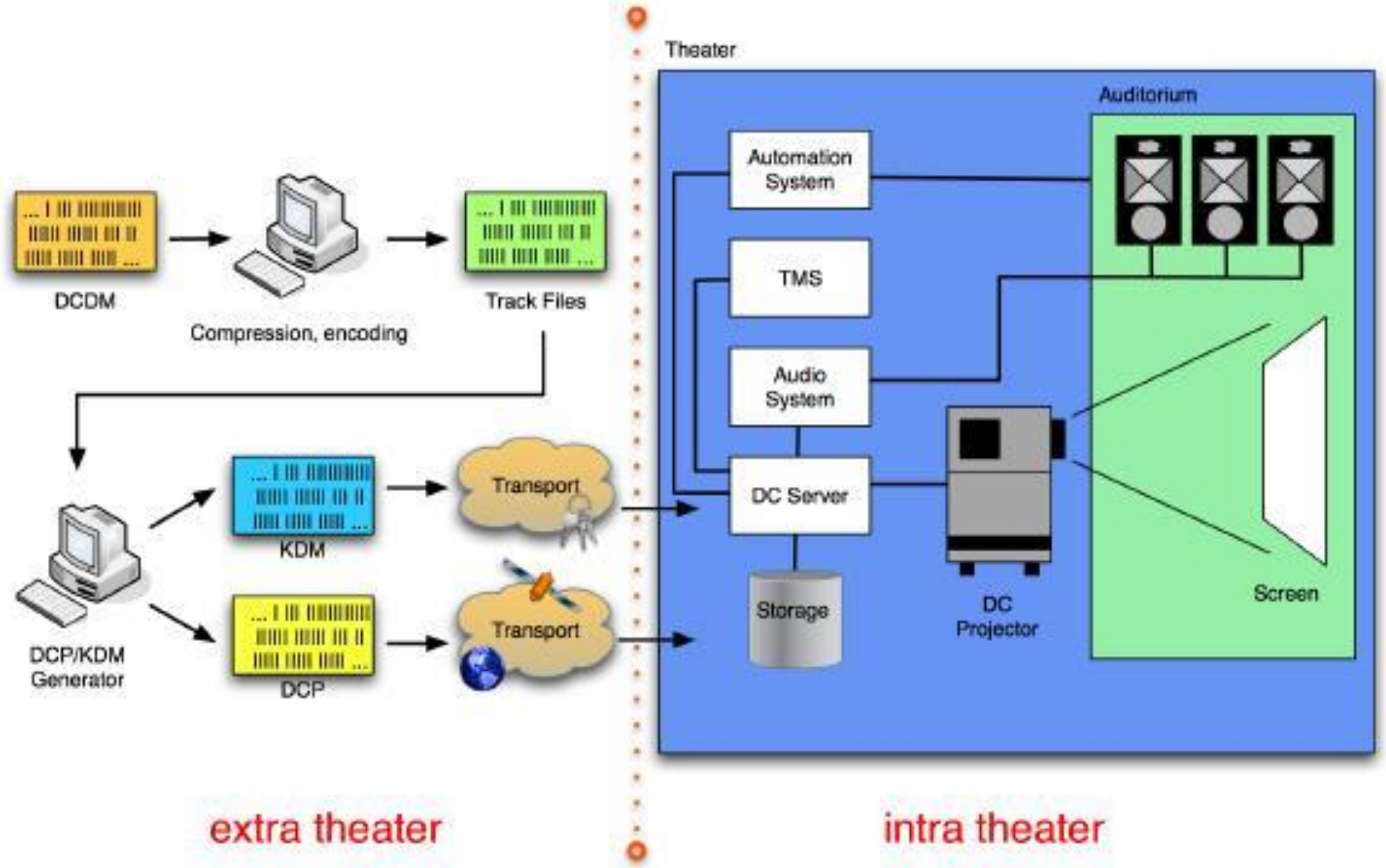


La Cadena del Cine

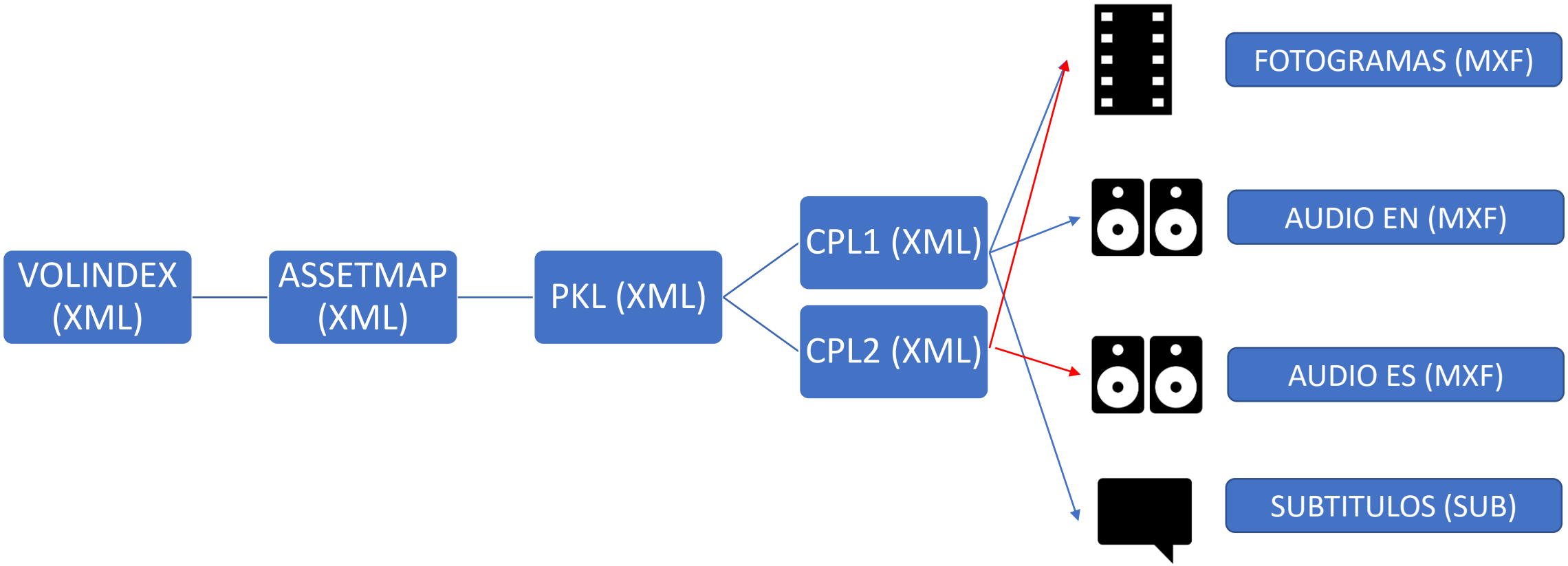


Edwards and Foessel, 2001 (http://slideplayer.com/slide/6034688)

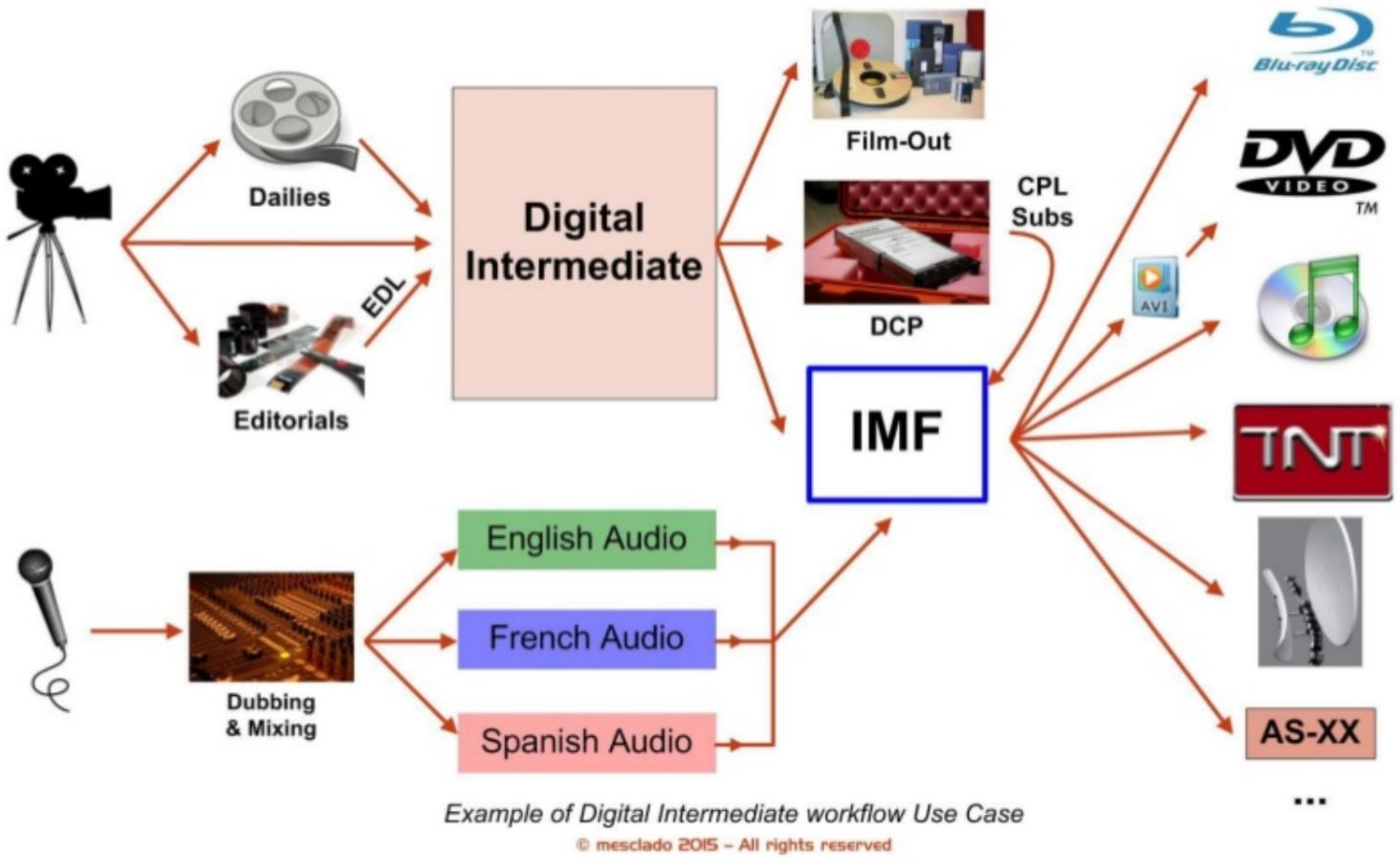
Sistema estandarizado DCI



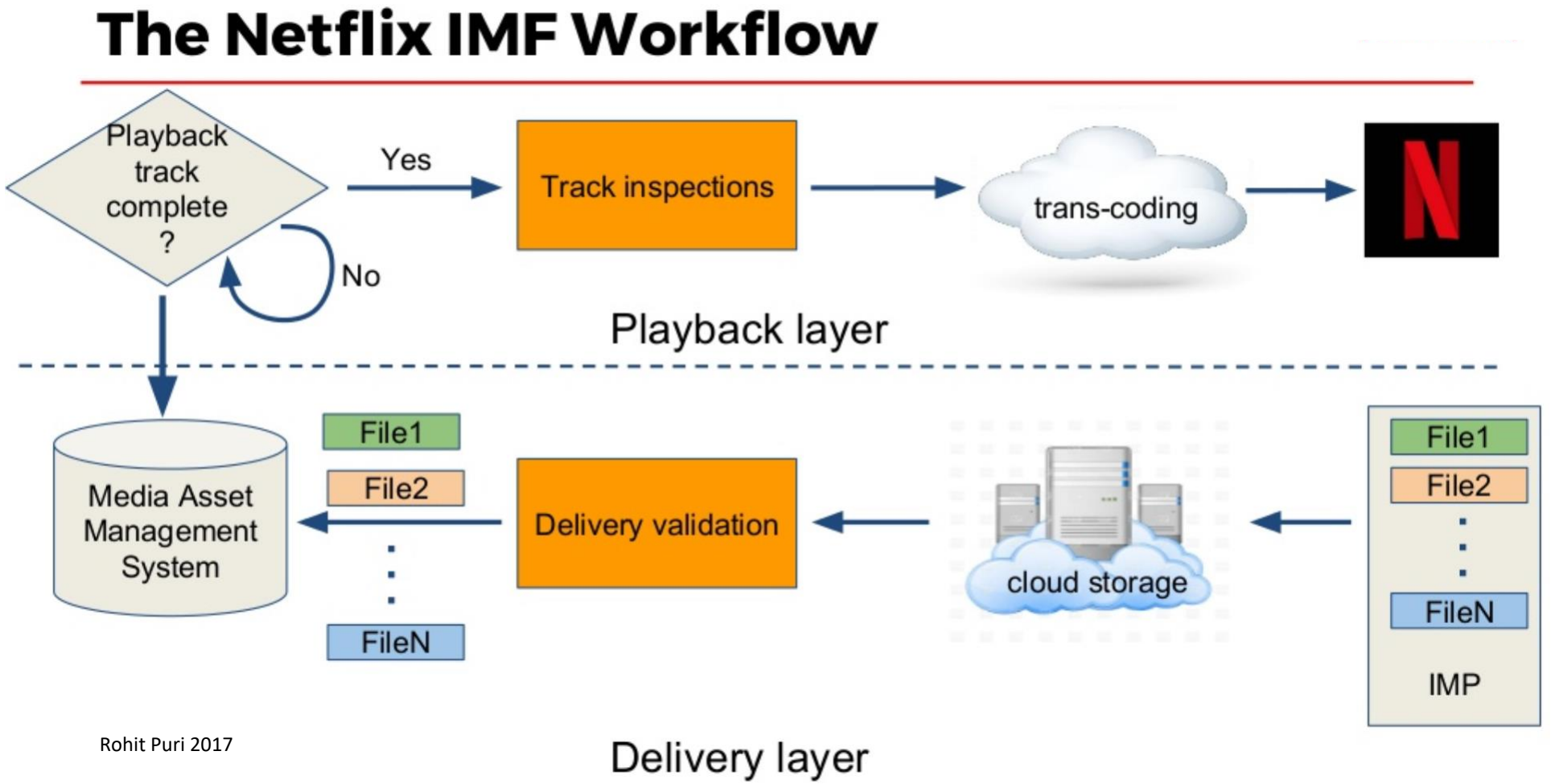
Estructura de un DCP



Más allá del cine: IMF



IMF para intercambio y almacén de contenido



Rohit Puri 2017

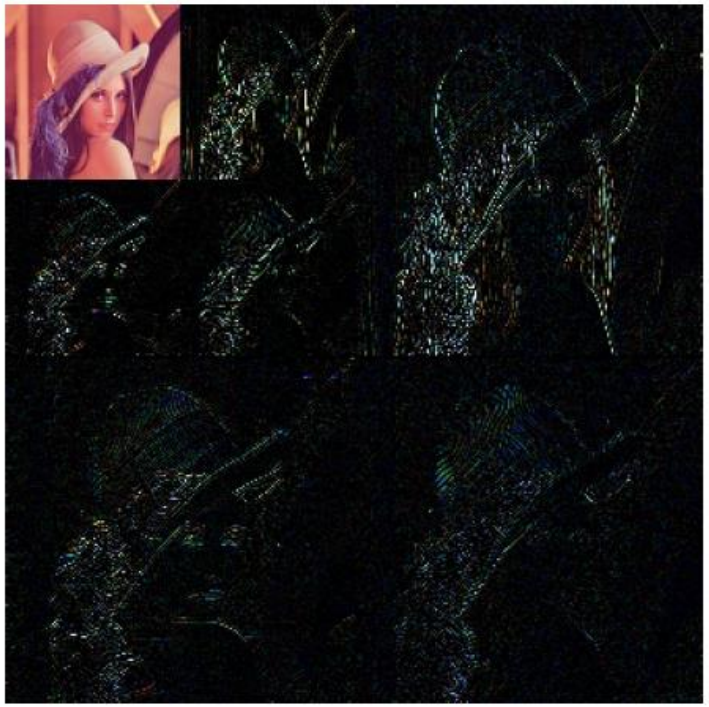
La base: el formato JPEG2000

- Perfiles: **Lossy** y **Lossless**
- Escalable: de una imagen se puede extraer “independientemente”
  - *Resolución*
  - *Calidad*
  - *Componente*
- **Intra-compression**: en vídeo, cada fotograma se comprime con J2K
- Menor pérdida de calidad visual comparado con JPEG
- Necesita **alta** carga computacional

La base: el formato JPEG2000

- Basado en transformada de Wavelet





Agenda

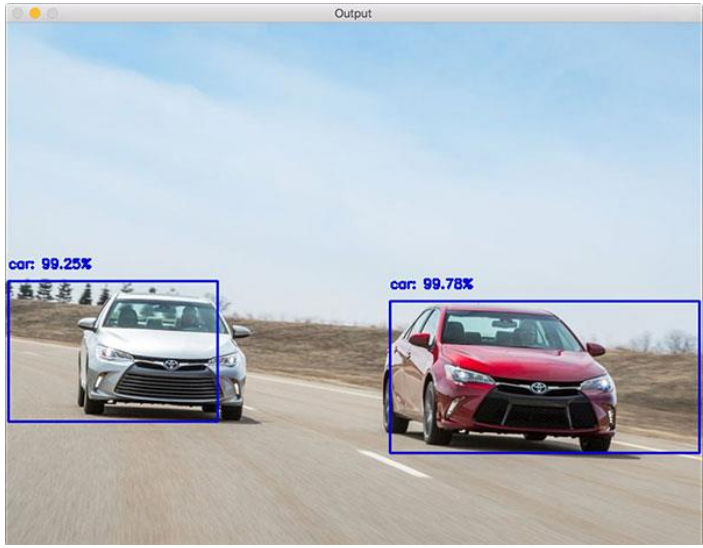
- Transfer Learning (conceptos básicos)
- Cine Digital
- Casos de estudio:
  - **Tag extraction**
  - Beat Event Detection

Tag extraction
Introducción

- Trabajo publicado en:
  - U.A. Kahn, N. Ejaz, M.A. Martínez-del-Amor, H. Sparenberg. **Movies Tags Extraction Using Deep Learning**. *14th IEEE International Conference on Advanced Video and Signal Surveillance (AVSS 2017), Lecce, Italy, 29 August - 1 September 2017.* Proceedings (October 2017), IEEE Xplore, pp. 1-6.
- Enlace al proyecto:
  - http://umair-khan.quest.edu.pk/announcements

Tag extraction
(extracción de etiquetas)

- **Extracción** (automática) de etiquetas de películas:
  - *Metadatos imprecisos* generado por humanos
  - Extracción de *información saliente* usando machine learning
  - *Semántica* de “alto nivel”
- **Idea**: selección de etiquetas clave, representando el tema general
- **Diferente** a la mayoría de tareas de reconocimiento de objetos o escenas:
  - No nos interesa si en un vídeo aparece una escopeta
  - Nos interesa saber si el video es de violencia, acción, etc.

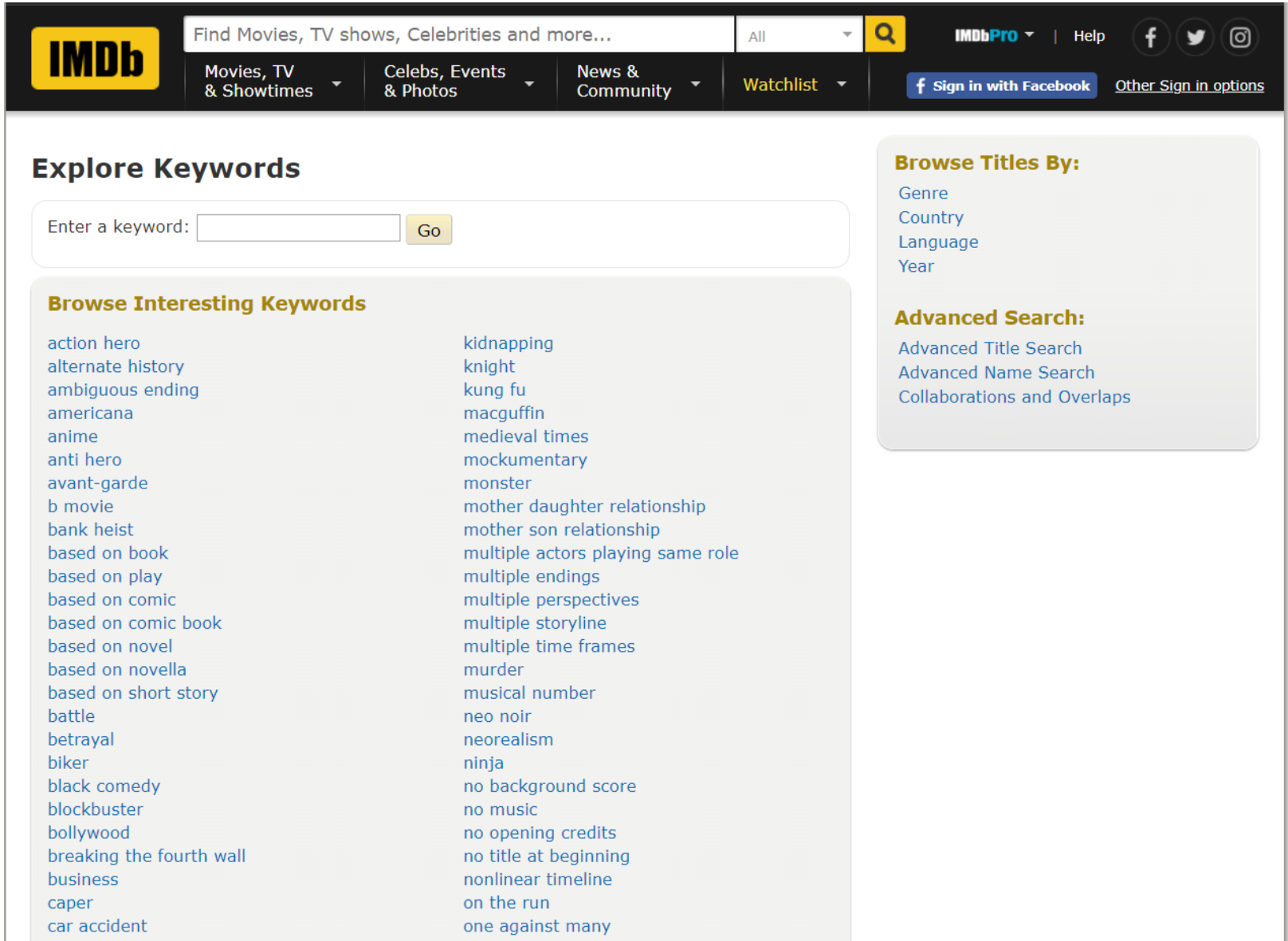


Object detection: 2 cars
vs
Movie tag: car chase

Tag extraction

Introducción

- Etiquetas ≈ keywords de IMDb
- Haremos solo una selección de ellas



Tag extraction
Introducción

- Aplicaciones potenciales:
  - Búsqueda por consulta
  - Almacenamiento y clasificación eficiente
  - Censura de contenido (violencia, desnudez, etc)
  - Sistemas de recomendación
  - Recuperación guiada por la escena
  - Asistencia en traducción de películas a lenguaje natural
  - Reconocimiento de acciones y comportamiento

Tag extraction
Diseño conceptual

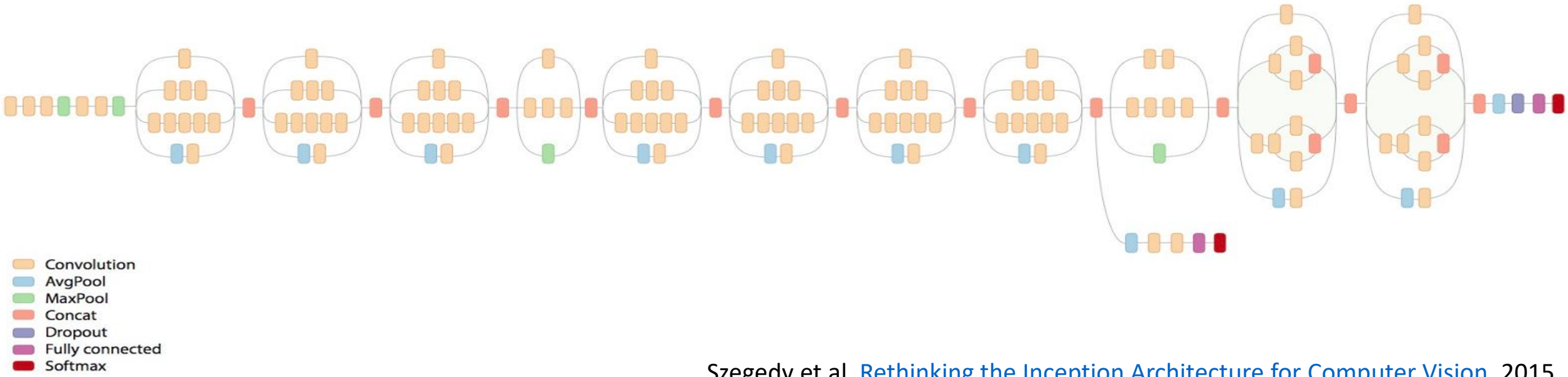
- **Problemática**:
  - No existe un dataset para etiquetas
  - Confección de uno desde cero y de forma manual
  - No disponíamos de recursos computacionales ni de mucho tiempo (beca posdoctoral).

- **Solución**: Transfer Learning para Fixed Feature Extractor
  - Necesitamos un dataset “mediano”
  - Inception-v3: Suficiente para nuestros PCs

Tag extraction

Elección modelo pre-entrenado

- **Inception-v3**:
  - Entrenado sobre ImageNet 2012 y 1000 clases
  - 3.46% tasa error
  - 48 capas
  - Disponible en tensorflow y tutoriales

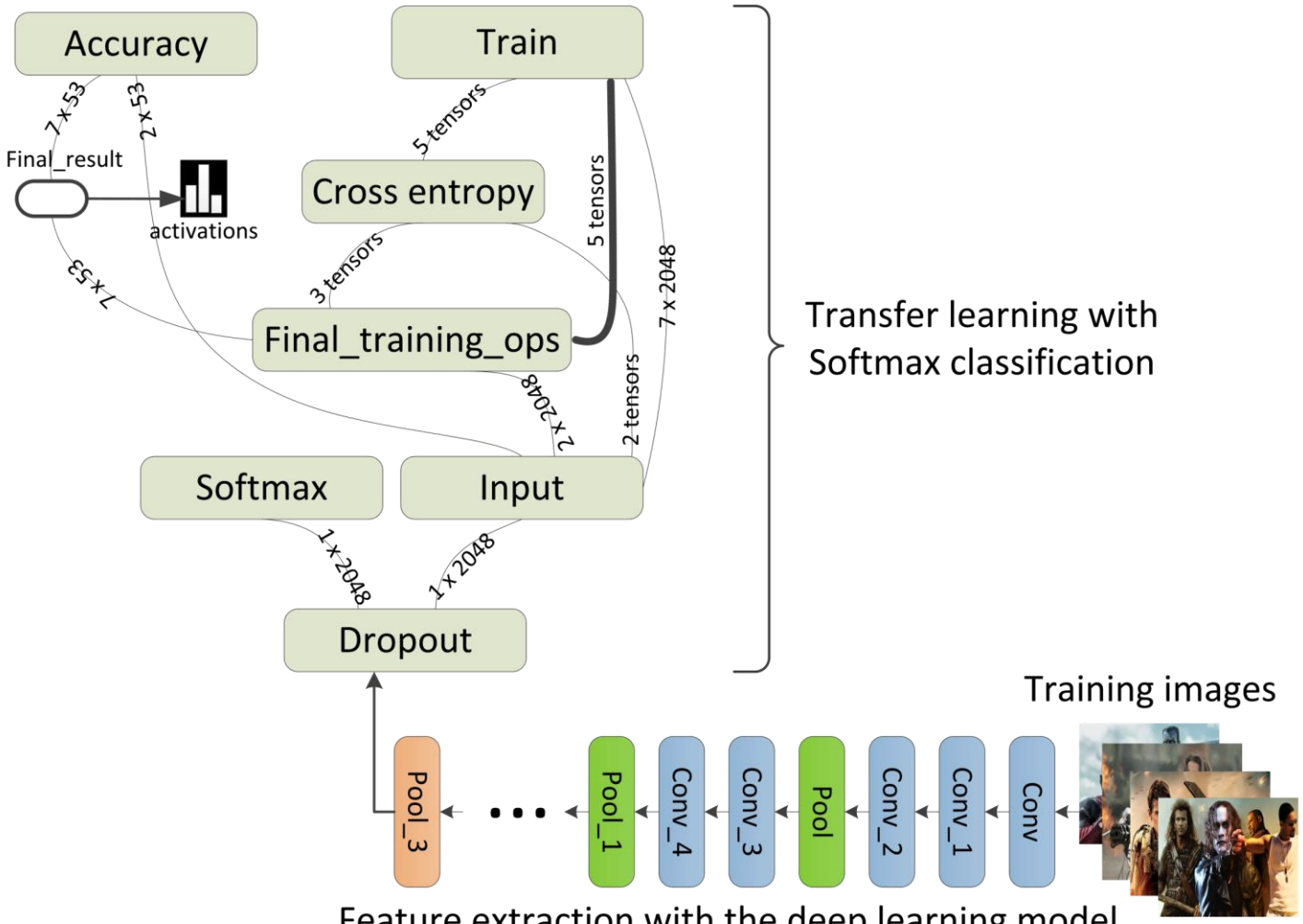


Szegedy et al. Rethinking the Inception Architecture for Computer Vision. 2015

Tag extraction

Transfer learning sobre Inception-v3

- Eliminada última capa
- Añadida una capa **Dropout**:
  - Desecha 50% activaciones aleatoriamente
  - Evita overfitting
- Después, activación **ReLU**:
  - No linearidad
  - $y_i = ReLU[\sum_j W_{i,j} x_j + b_i]$
- Clasificación con **Softmax**:
  - Convertir a probabilidad
  - $p_i = \frac{e^{y_i}}{\sum_j^{50} e^{y_j}}$



Tag extraction

Conjunto de datos

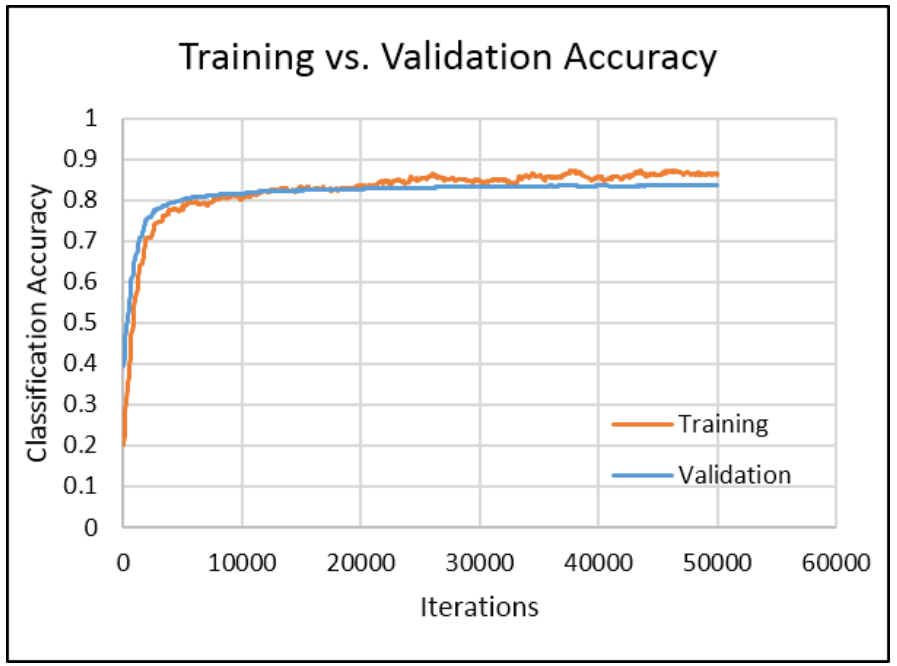
- Vocabulario 50 etiquetas (con solapamiento)
- 700 imágenes/etiqueta

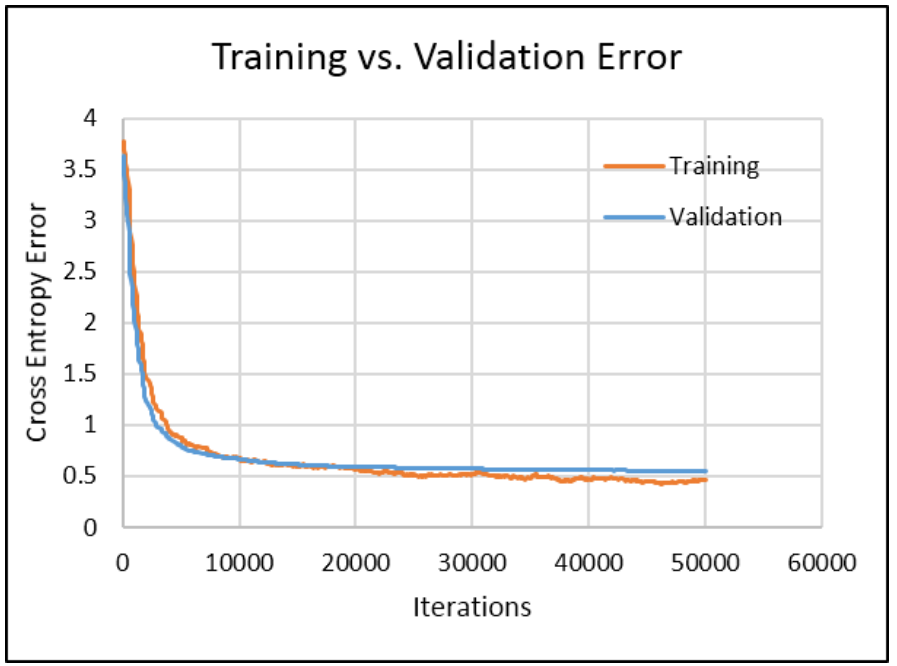
Action	Bomb explosion	Car chase
Destruction	Sword fight	Vehicle crash
Violence	Abduction	Heist
Adventure	Animal	Beach/Sea
Climbing	Desert	Hiking
Forest	Valleys/Hills	Children
Family	Club/Bar	Dance
Music	Wedding	College/Univ.
Hospital	Drinking	Food
Smoking	Exercise	Sports
Swimming	Glamor/Fashion	Nudity
Romance	Sex	Horror
Monster	Murder	Lab Experiment
Sci-fi	Super hero	Technology
Robot	Military	Police
Prison	War	Weapon
Animation	Drama	

Tag extraction

Resultados en fotogramas individuales

- **Distribución**: 80% entrenamiento, 10% validación y 10% test
- **Cross entropy**, función de perdida para estimar error: $E(p,q) = -\sum_{x}^{50} q(x) \log p(x)$
- **500 epochs**





Tag extraction

Resultados en fotogramas individuales

Military, action, weapon, war

Violence, destruction, bomb explosion, action, car crash

Sex, nudity, romance, modeling

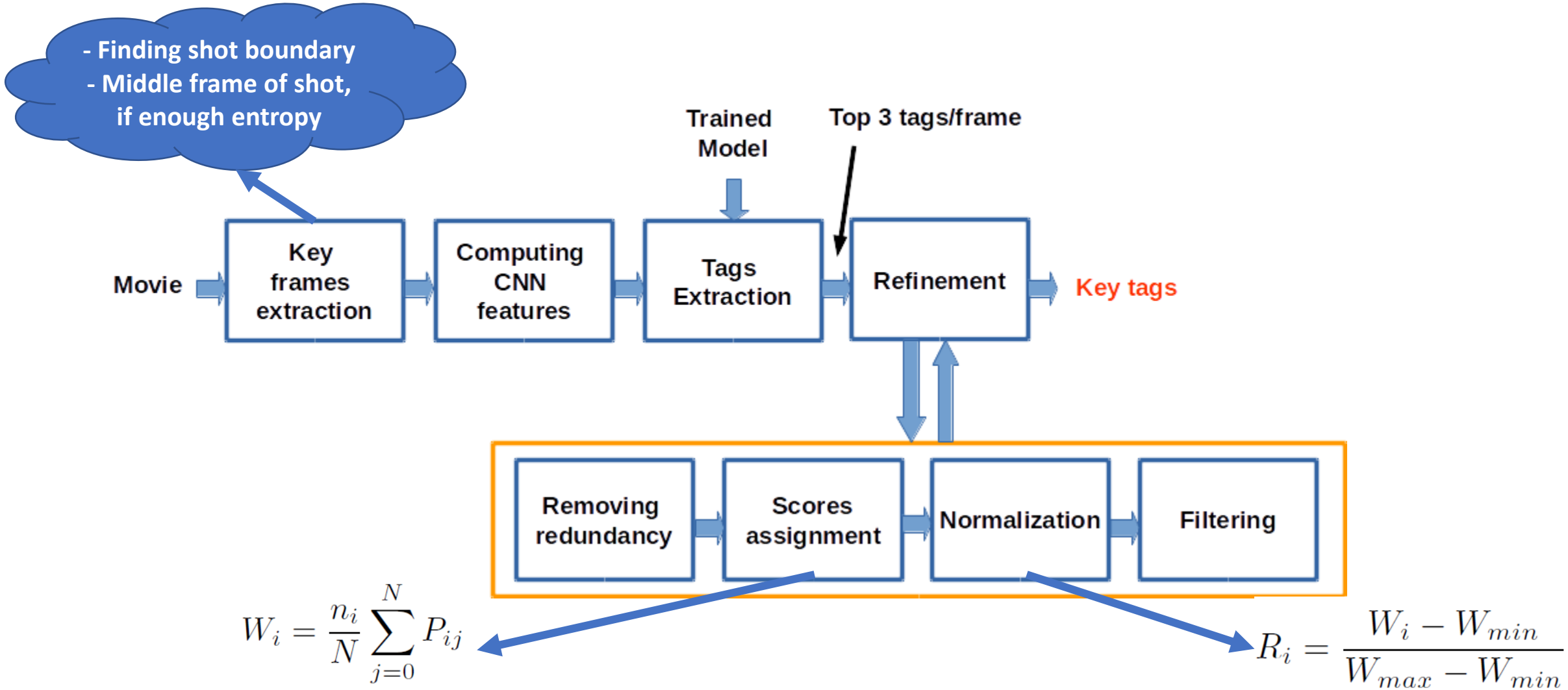
Hiking, adventure, nature, forest, valleys, hills, climbing

Sci-fi, super hero, robot, action

Violence, sci-fi, action, horror

Tag extraction

Extrapolación a vídeo



Tag extraction
Experimentación

- Configuración Hardware/Software:

Hardware/Software	**Specifications**
CPU	Intel Xeon(R) E5430, 2.66GHz x 8
RAM	8GB
GPU	GeForce GTX 1050 Ti, 768 cores, 4GB GDDR5
Deep learning framework	Tensorflow 1.0, compiled with GPU support
Operating System	Ubuntu 16.04 (64-bit)
Programming languages	Python 2.7, OpenCV 3.0, C++

Tag extraction
Experimentación

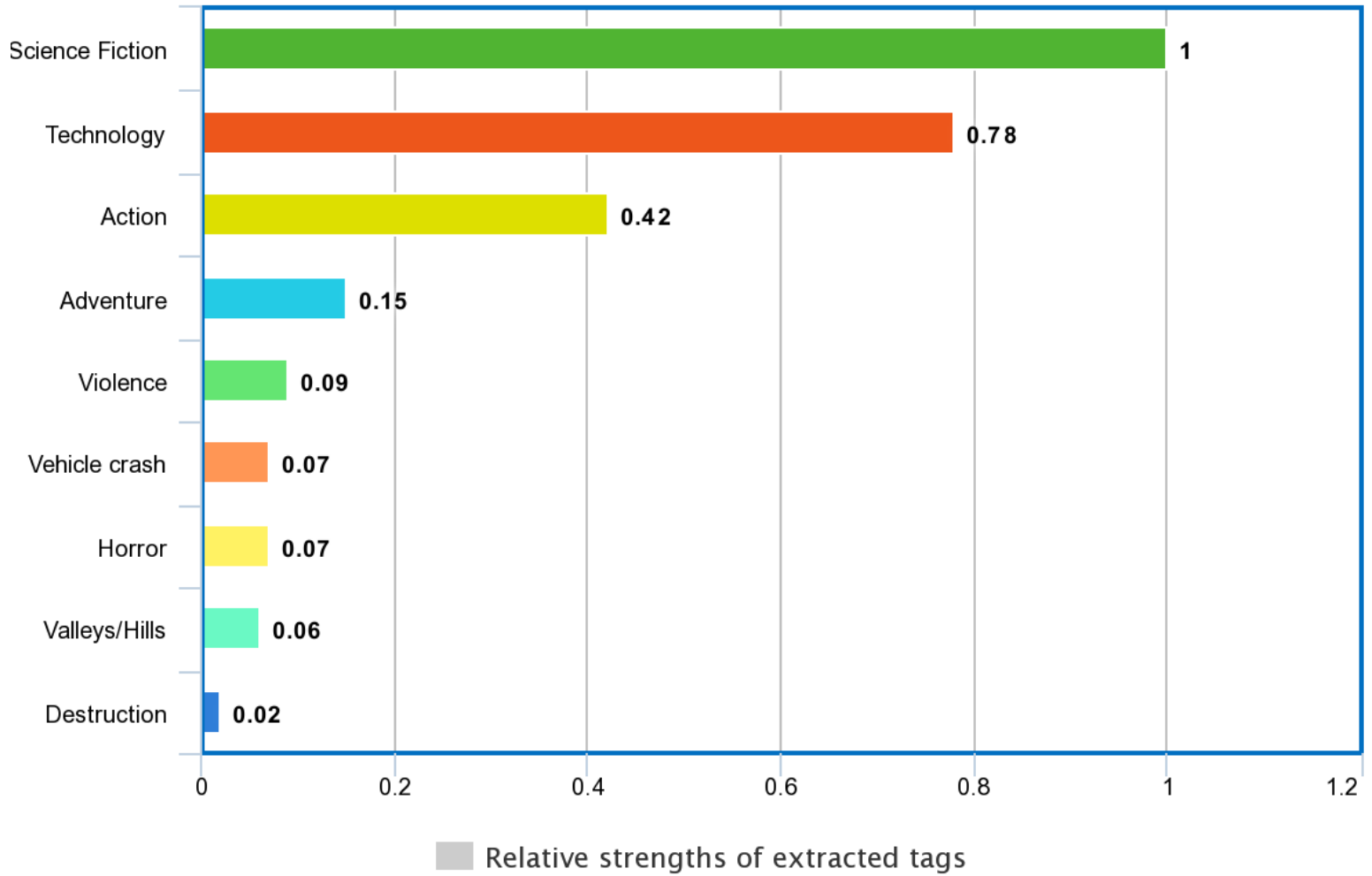
- Problemática:
  - **No hay un ground truth**, o marco de referencia
- Realización de **3 experimentos subjetivos:**
  - Llevados a cabo en el cine del Fraunhofer IIS
  - Muestra de 10 tráilers de películas
  - 10 voluntarios distintos en cada uno
- **Medidas**:
  - Mean Opinion Score: *de una encuesta, cuantos resultados corresponden*
  - Precisión: *cuántos de los resultados positivos son correctos*
  - Recall: *cuántos resultados positivos correctos respecto a todos los positivos*
  - F-score: *media ponderada de precisión y recall*

Tag extraction
Experimentación

- **Experimento 1: tags rating**
  - Reparto de las etiquetas extraídas por nuestro algoritmo para cada vídeo
  - Voluntarios valora cada uno por separado.
  - Mean Opinion Score: 84.3%
  - Ejemplo: RAW (2017) tráiler
    - Romance, violence, action, car crash, horror, sex, child, nudity, outdoor/nature/forest, hospital, food, Club/bar, college/university, music, crowd

Tag extraction
Experimentación

- **Experimento 2: tags rating w.r.t. relevancy and strength**
  - Igual que experimento 1, pero teniendo en cuenta la relevancia de cada etiqueta.
  - Mean Opinion Score: 77.8%
  - Ejemplo: Alien Covenant teaser trailer



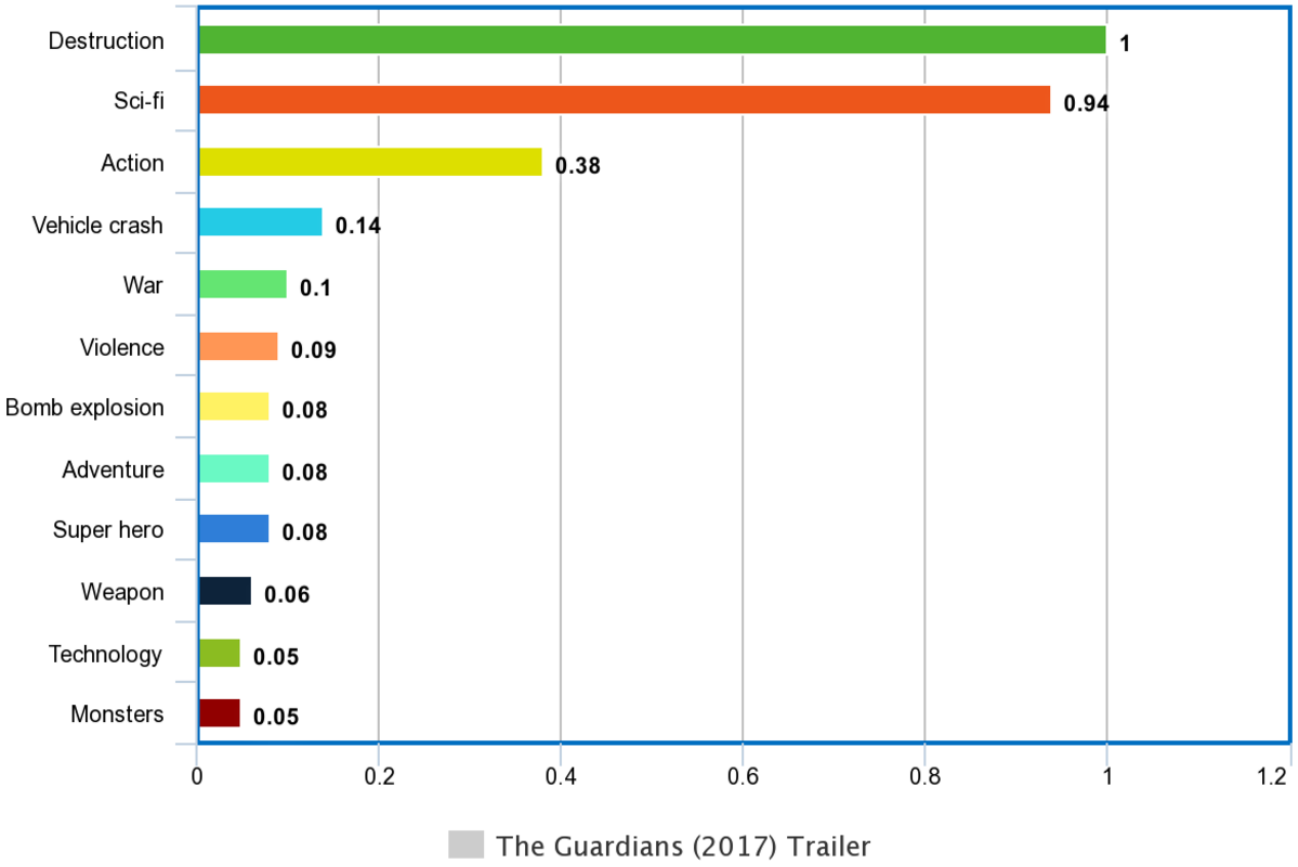
Tag extraction
Experimentación

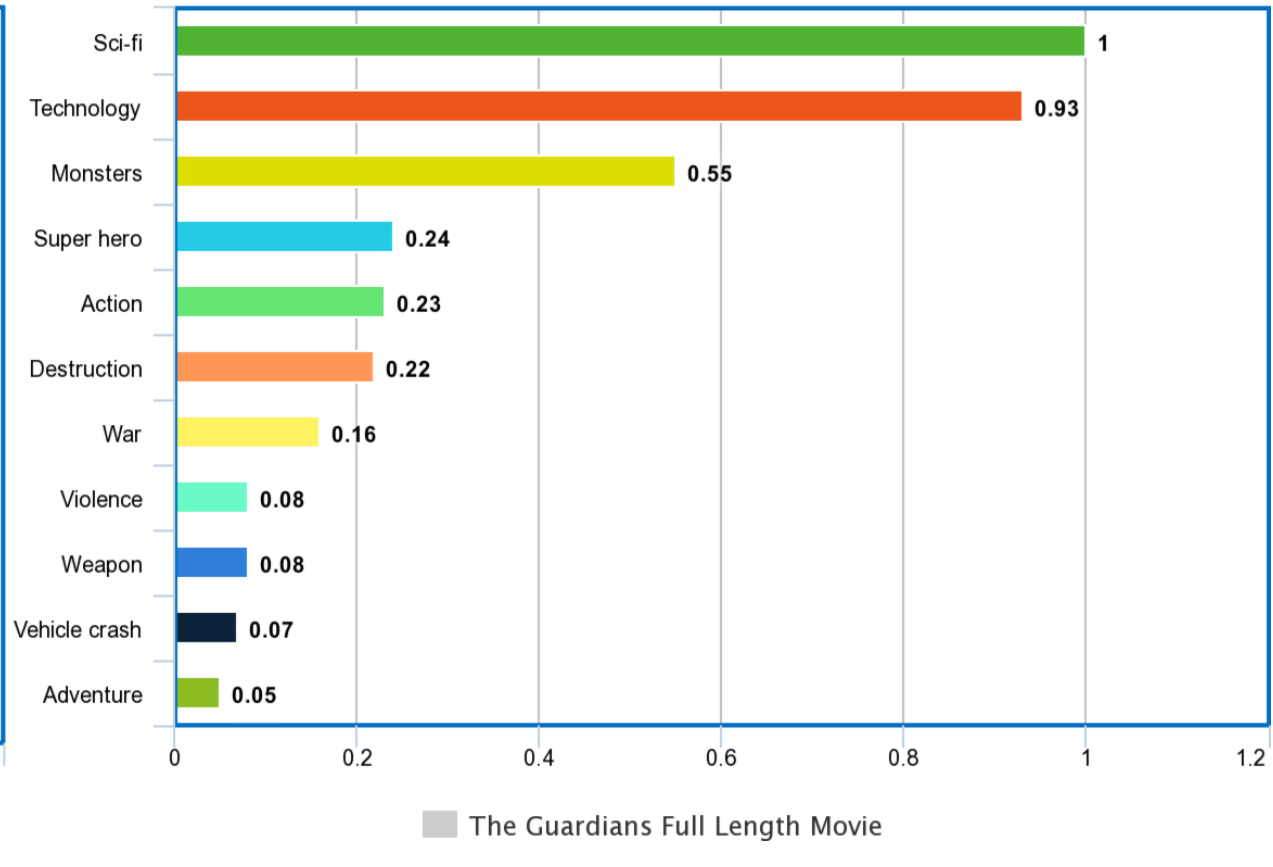
- **Experimento 3: tags matching**
  - Repartir todo el vocabulario de etiquetas, repetido para cada vídeo
  - Voluntarios eligen las etiquetas que crean más relevantes
  - Usando este experimento como ground truth:
    - MAP = 76%, MAR = 74.22%
    - F1 - Score = 0.75%

Tag extraction

Experimentación (tráiler vs película completa)

- **Trailer duration:** 2 min, 26 sec, **Processing time:** 17sec
- **Full length movie duration:** 1 hr, 24 min, **Processing time:** 10 min

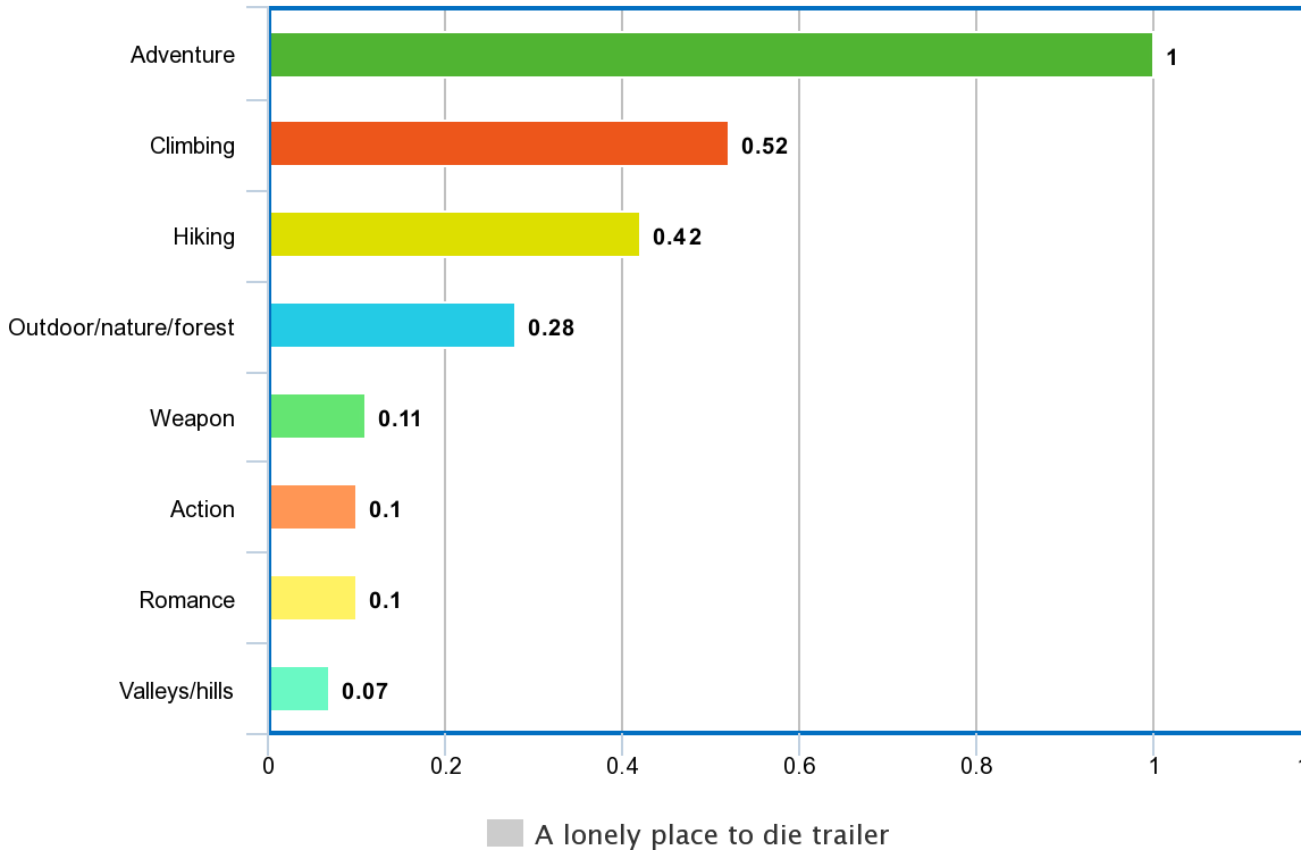


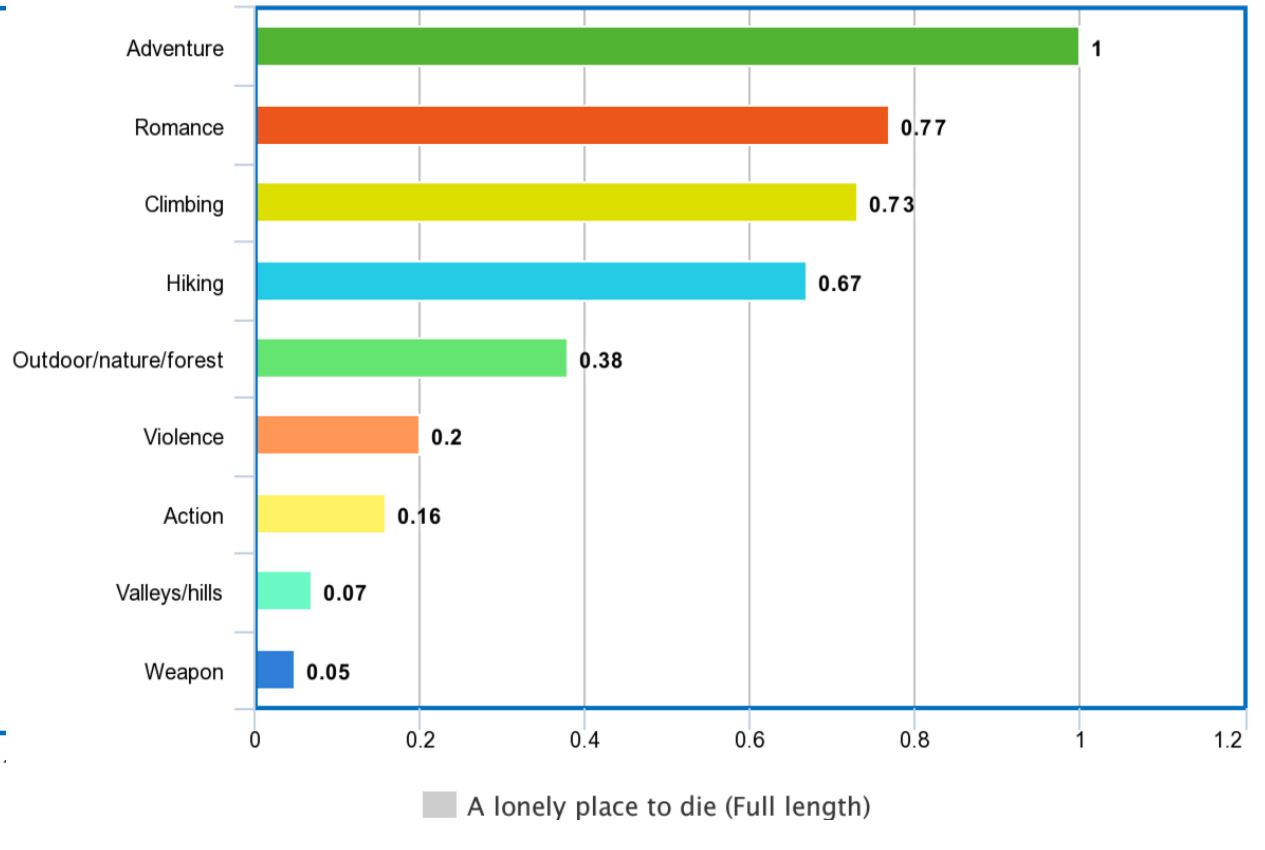


Tag extraction

Experimentación (tráiler vs película completa)

- **Trailer duration:** 1 min, 53 sec, **Processing time:** 13 sec
- **Full length movie duration:** 1 hr, 39 min, **Processing time:** 7 min





Tag extraction
Trabajo futuro

- Extensión del vocabulario de etiquetas
- Aprender la correlación semántica entre etiquetas
- Extracción de los fotogramas clave más relevantes para resumen de películas
- Extracción de escenas basado en consulta

Tag extraction

Extra: video segmentation

- Trabajo aún sin publicar.
- Creación de vídeo conteniendo escenas de solo una selección de etiquetas.
- Ver tráiler y ejemplo.

Agenda

- Transfer Learning (conceptos básicos)
- Cine Digital
- Casos de estudio:
  - Tag extraction
  - **Beat Event Detection**

Beat Event Detection

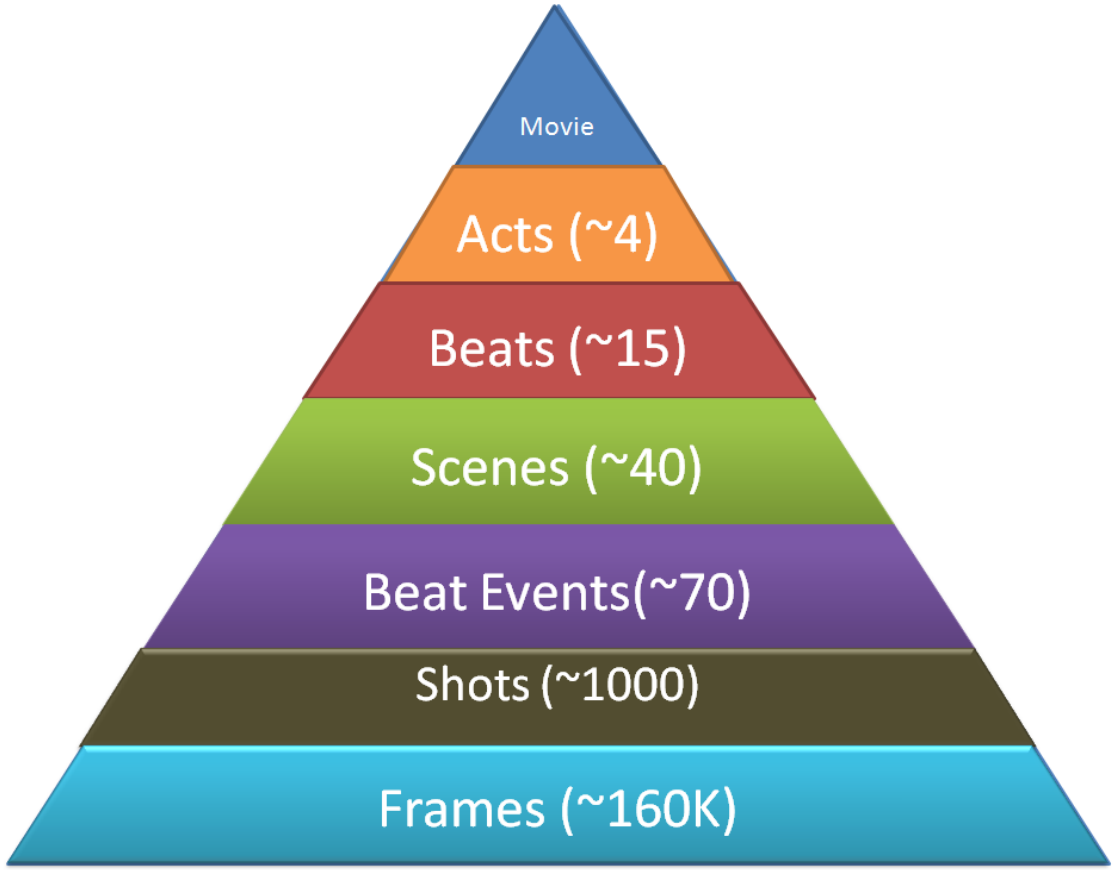
(detección de eventos de ritmo)

- Trabajo publicado en:
  - N. Ejaz, U.A. Kahn, M.A. Martínez-del-Amor, H. Sparenberg. **Deep Learning based Beat Event Detection in Action Movie Franchises**. *The 10th International Conference on Machine Vision (ICMV 2017), Vienna, Austria, November 13-15, 2017.* Proceedings (April 2018), SPIE 10696, 1069608.
- Enlace al proyecto:
  - http://www.naveedejaz.net/beat-even-detection-in-movies.html

Beat Event Detection

Introducción

- Una película es temporalmente subdividida en una jerarquía de actos, escenas, planos y fotogramas.
- Los “beats” son cambios en la historia de una película, dando ritmo.
- “*Beat events*” son una consecución de planos con una actividad común.
- **Objetivo**: extraer las delimintaciones de los beat events y asignarle una etiqueta.



Beat Event Detection

Action Movie Franchise dataset

- Basado en el dataset de beat events Action Movie Franchise.

Movie Franchises	Movie Names
RAMBO	First Blood(1982) First Blood II(1985) Rambo III(1988) Rambo IV(2008)
ROCKY	Rocky(1976) Rocky II(1979) Rocky III(1982) Rocky IV(1985)
INDIANA JONES	Raiders of the Lost Ark (1981) Indiana Jones and the Temple of Doom (1984) Indiana Jones and the Last Crusade (1989) Indiana Jones and the Kingdom of the Crystal Skull (2008)
LETHAL WEAPONS	Lethal Weapon(1987) Lethal Weapon 2 (1989) Lethal Weapon 3 (1992) Lethal Weapon 4 (1998)
DIE HARD	Die Hard (1988) Die Hard II(1990) Die Hard with a Vengeance (1995) Live Free or Die Hard(2007)

Potapov et al. Beat-event detection in action movie franchises. 2015. CoRR, abs/1508.03755.

Beat Event Detection

Action Movie Franchise dataset

- Realizado en INRIA
- **Anotación** de shots y beat events de 20 películas
- **11** categorías
- Cobertura 60% (etiqueta *Difficult* cuando ambiguo)

No.	Beat Event Category
1	Pursuit
2	Battle Preparation
3	Battle
4	Romance
5	Despair Good
6	Joy Bad
7	Good Argue Good
8	Good Argue Bad
9	Bad Argue Bad
10	Victory Good
11	Victory Bad

Potapov et al. Beat-event detection in action movie franchises. 2015. CoRR, abs/1508.03755.

Beat Event Detection

Action Movie Franchise dataset

- Ejemplos de fotogramas con categorías del beat event correspondiente.

Potapov et al. Beat-event detection in action movie franchises. 2015. CoRR, abs/1508.03755.

Beat Event Detection

Action Movie Franchise dataset

- **Trabajo previo** por Potapov et al.:
  - Computación de **descriptores** de shots, sobre canales de audio y visual, usando:
    - Dense SIFT
    - CNN features
    - Motion descriptors
    - Audio descriptors
    - Face descriptors
  - Clasificación de shots con **SVM**.
  - Shots vecinos agrupados para clasificar beat events.

- **Aplicaciones potenciales**, mismas que con tag extraction, además:
  - Además: compression a nivel de escena y beat events en streaming.

Beat Event Detection

Diseño conceptual

- Aproximación con DL:
  - Construcción de **dataset propio** (750 imágenes/categoría)
  - Diferentes fuentes: Youtube, trailers, otras películas

- **Transfer Learning**, con configuración similar a Tag Extraction:
  - Inception-V3
  - Eliminación capa clasificación, y añadido de dropout y softmax
  - 500 epochs
  - Test accuracy del 85%

Beat Event Detection

Diseño conceptual

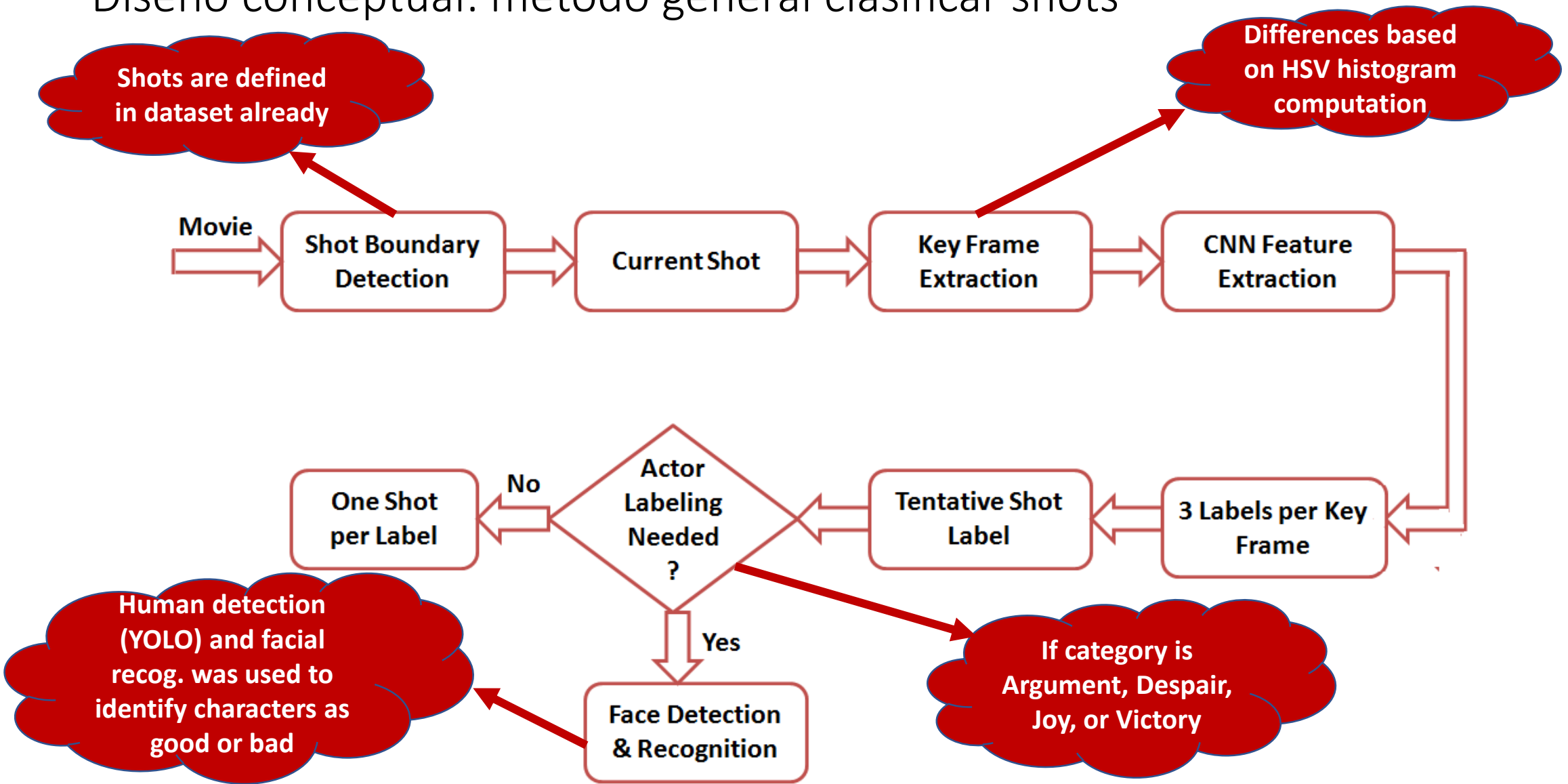
- Modificación de las categorías para mejorar precisión
  - **Partición** de categorías en subcategorías

Category	Sub-Categories
Battle	Physical War, War Destruction, Bomb Explosion, Fire Attack, Arrows' War
Pursuit	Car Pursuit, Bike Pursuit, Horse Pursuit, Men Pursuit
Despair	Despair by Expressions, Despair in Jail

  - **Unión** de categorías (clasificación en dos etapas)
    - *Good Argue Good, Good Argue Bad, Bad Argue Bad → Argument*
    - *Joy Bad → Joy*
    - *Despair Good → Despair*
    - *Victory Good and Victory Bad → Victory*

Beat Event Detection

Diseño conceptual: método general clasificar shots



Beat Event Detection

Diseño conceptual: método general clasificar beat events

- **Ventana deslizante** sobre las etiquetas de cada shot (tamaño 3)
- **Dos casos**:
  - Todas etiquetas son distintas → Cada shot es un beat event
  - No todas las etiquetas son distintas → Elegir etiqueta mayoritaria
- **Extender el beat event** si en las siguientes iteraciones sale misma etiqueta.

Beat Event Detection

Experimentación

- Configuración Hardware/Software

Hardware/Software	**Specifications**
CPU	Intel Xeon(R) E5430. 2.66GHz x 8
RAM	8GB
GPU	GeForce GTX 1050 Ti, 768 cores, 4GB GDDR5
DL Framwork	Tensorflow 0.12, compiled with GPU support
Operating System	Ubuntu 16.04 (64 bits)
Programming Languages	Python 2.7, OpenCV 3.0

Beat Event Detection

Experimentación: ejemplos

Battle

Good Argue Good

Pursuit

Romance

Beat Event Detection

Experimentación: resultados

Beat Event	Precision*	Precision	Recall	F-measure
Battle	0.39	0.64	0.91	0.76
Good Argue Good	0.17	0.81	0.81	0.81
Good Argue Bad	0.12	0.59	0.86	0.70
Romance	0.23	0.69	0.70	0.70
Despair Good	0.06	0.89	0.49	0.63
Preparation	0.27	0.86	0.06	0.10
Pursuit	0.35	0.63	0.82	0.71
Joy Bad	0.05	0.58	0.71	0.64
Bad Argue Bad	0.06	0.58	0.58	0.58
Victory Bad	0.04	1.00	0.33	0.50
Victory Good	0.15	0.92	0.26	0.41
Mean	0.17	0.74	0.59	0.59

Beat Event Detection

Experimentación: matriz de confusión

	Batt	GAG	GAB	Rom	DG	Prep	Purs	JB	BAB	VB	VG
Batt	354	2	1	4	3	0	22	2	0	0	0
GAG	5	237	10	13	4	0	18	5	0	0	0
GAB	3	0	120	1	3	0	5	6	1	0	0
Rom	1	19	0	69	3	0	2	2	2	0	0
DG	66	10	33	8	186	0	35	40	3	0	0
Prep	47	23	4	3	5	6	16	3	1	0	1
Purs	30	0	5	0	2	0	168	0	0	0	0
JB	3	0	26	1	1	0	0	84	3	0	0
BAB	6	0	1	1	1	0	1	0	14	0	0
VB	6	0	2	0	1	1	0	2	0	6	0
VG	28	2	3	0	0	0	1	0	0	0	12

* Battle=Batt, Good Argue Good= GAG, Romance=Rom, Despair Good=DG, Preparation= Prep, Pursuit=Purs, Joy Bad=JB, Bad Argue Bad= BAB, Victory Bad= VB, Victory Good= VG

Beat Event Detection

Trabajo futuro

- Incluir información de otras fuentes multimedia:
  - Audio
  - Subtítulos
- Otros tipos de películas (no de acción)
- Aplicaciones

Muchas gracias

¿Preguntas?